NAVAL POSTGRADUATE SCHOOL

Monterey, California

DISSERTATION

AN INVESTIGATION OF
MULTIVARIATE ADAPTIVE REGRESSION SPLINES
FOR MODELING AND ANALYSIS
OF UNIVARIATE AND SEMI-MULTIVARIATE
TIME SERIES SYSTEMS

by

James G. Stevens

September 1991

Thesis Advisor: Peter A. W. Lewis

Unclassified
ecurity Classification of this Page

REPORT DOCUMENTATION PAGE

a. Report Security Classification: Unclassified	1b. Restrictive Markings
a. Security Classification Authority	3. Distribution/Availability of Report: Approved for public release; distribution is unlimited.
b. Declassification/Downgrading Schedule	
Performing Organization Report Number(s)	5. Monitoring Organization Report Number(s)
a. Name of Performing Organization: Naval Postgraduate School / 6b. Office Symbol: OR	7a. Name of Monitoring Organization: Naval Postgraduate School
c. Address (*City, State, and ZIP code*): Monterey, CA 93943-5000	7b. Address (*City, State, and ZIP code*): Monterey, CA 93943-5000
a. Name of Funding/Sponsoring Organization / 8b. Office Symbol	9. Procurement Instrument Identification Number
c. Address (*City, State, and ZIP code*)	10. Source of Funding Numbers: Program Element No / Project No / Task No / Work Unit Accession No

1. Title (*Include Security Classification*)
AN INVESTIGATION OF MULTIVARIATE ADAPTIVE REGRESSION SPLINES FOR MODELING AND ANALYSIS OF UNIVARIATE AND SEMI-MULTIVARIATE TIME SERIES SYSTEMS

2. Personal Author(s): James G. Stevens

3a. Type of Report	13b. Time covered	14. Date of Report (*year, month, day*)	15. Page Count
Ph.D. Dissertation	From To	Sept 1991	207

6. Supplementary Notation: The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

17. Cosati Codes			18. Subject Terms (*continue on reverse if necessary and identify by block number*)
eld	Group	Subgroup	Adaptive Regression Splines, Threshold Autoregression, Times Series Systems, Simulation, Model Selection

9. Abstract (*continue on reverse if necessary and identify by block number*)

This dissertation investigates the use of multivariate adaptive regression splines (MARS), due to Friedman, for nonlinear regression modeling and analysis of time series systems. MARS is a computationally intensive methodology that fits a nonparametric regression model in the form of an expansion in product spline basis functions of predictor variables chosen during a forward and backward recursive partitioning strategy. The MARS algorithm produces continuous nonlinear regression models for high-dimensional data using a combination of predictor variable interactions and partitions of the predictor variable space. By letting the predictor variables in the MARS algorithm be lagged values of a time series system, one obtains a univariate (ASTAR) or semi-multivariate (SMASTAR) adaptive spline threshold autoregressive model for nonlinear autoregressive threshold modeling and analysis of time series, thereby extending the threshold autoregression (TAR) time series methodology developed by Tong. The models seem well suited for taking into account the complex interactions among multivariate, cross-correlated, lagged predictor variables of a time series system. A difficulty faced during regression modeling is the problem of model selection. Currently, a modified form of generalized cross validation (GCV^*) is used for model selection within the MARS algorithm. However, one question that immediately develops is whether GCV^* is the 'best' criterion for model selection when using serially and cross-correlated time series data. Using MSE as a measure of performance, simulations show that other model selection criteria, in particular the Schwarz-Rissanen (SC) criterion, can improve model selection over GCV^*.

0. Distribution/Availability of Abstract	21. Abstract Security Classification	
☒ unclassified/unlimited ☐ same as report ☐ DTIC users	Unclassified	
2a. Name of Responsible Individual: Peter A. W. Lewis	22b. Telephone (*include Area code*): (408) 646-2283	22c. Office Symbol: ORLw

D FORM 1473, 84 MAR — Previous editions are obsolete. — Security Classification of this Page: Unclassified

Approved for public release; distribution is unlimited

An Investigation of Multivariate Adaptive Regression Splines for Modeling and Analysis of Univariate and Semi-Multivariate Time Series Systems

by

James G. Stevens

Major, United States Army

B. S., United States Military Academy, 1977

M.S., Naval Postgraduate School, 1987

Submitted in partial fulfillment of the

requirements for the degree of

DOCTOR OF PHILOSOPHY IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL

September, 1991

ABSTRACT

This dissertation investigates the use of multivariate adaptive regression splines (MARS), due to Friedman, for nonlinear regression modeling and analysis of time series systems. MARS can be conceptualized as a generalization of recursive partitioning that uses spline fitting in lieu of other simple fitting functions. MARS is a computationally intensive methodology that fits a nonparametric regression model in the form of an expansion in product spline basis functions of predictor variables chosen during a forward and backward recursive partitioning strategy. The MARS algorithm produces continuous nonlinear regression models for high-dimensional data using a combination of predictor variable interactions and partitions of the predictor variable space.

By letting the predictor variables in the MARS algorithm be lagged values of a time series system, one obtains a univariate (ASTAR) or semi-multivariate (SMASTAR) adaptive spline threshold autoregressive model for nonlinear autoregressive threshold modeling and analysis of time series, thereby extending the threshold autoregression (TAR) time series methodology developed by Tong. The models seem well suited for taking into account the complex interactions among multivariate, cross-correlated, lagged predictor variables of a time series system. A significant feature of this time series application of MARS is its ability to produce models with limit cycles when modeling time series data that exhibit periodic behavior. In a physical context, limit cycles represent a stationary state of sustained oscillations.

A difficulty faced during regression modeling is the problem of model selection, i.e., choosing the appropriate model dimension and model predictor variables. Currently, a modified form of generalized cross validation (GCV^*), first suggested by Craven and Wahba, is used for model selection within the MARS algorithm. However, one question that immediately develops is whether GCV^* is the 'best' criterion for model selection when using serially and cross-correlated time series data. Using MSE as a measure of performance, simulations show that other model selection criteria, in particular the Schwarz-Rissanen (SC) criterion, can improve model selection over GCV^*.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENT

I thank Professor Jerome Friedman for supplying his MARS programs. The computing and graphics in this dissertation was done with the experimental APL package GRAFSTAT from IBM Research; I am grateful to Dr. Peter Welch for supplying GRAFSTAT.

To P.A.W. Lewis,

Thank you for your support, confidence and persistence.

This dissertation is dedicated to those who are my life;

my loving wife, Bet

and our children,

Kathryn,

Jimmy,

Megan, and

Patrick.

I. INTRODUCTION

Most research in, and applications of, time series modeling and analysis has been concerned with linear models. This is due to the maturity of the theory for linear time series, and the numerous studies and statistical packages that exist to facilitate the use of linear time series models. However, more frequently than not, nonlinear time dependent systems abound that are not adequately handled by linear models. For these systems we need to consider general classes of nonlinear models that readily adapt to the precise form of a nonlinear system of interest. This dissertation is an investigation of the use of multivariate adaptive regression splines for the systematic autoregressive modeling and analysis of nonlinear univariate and semi-multivariate time series systems. This chapter provides a brief introduction to regression modeling and multivariate adaptive regression spline modeling (MARS), briefly discusses the application of MARS to time series systems, identifies the contributions of this dissertation, and gives an outline of the chapters that follow.

A. BACKGROUND

Regression modeling is a popular statistical approach that serves as a basis for studying a system of interest. Regression modeling is used to develop a mathematical model of the relationships that exist between the dependent (output) variable and the independent (explanatory) variables of the system. Classical methods for developing the functional form of the regression model are based on previous knowledge of the system and on considerations such as smoothness and continuity of the output variable as a function of the explanatory (predictor) variables.

To provide a framework for a regression modeling methodology let y represent a single response variable that depends on a vector of p predictor variables $\boldsymbol{x}$, where $\boldsymbol{x} = (x_1, \ldots, x_v, \ldots, x_p)$. Assume there are given N samples of y and $\boldsymbol{x}$, namely $\{y_i, \boldsymbol{x}_i\}_{i=1}^{N}$, and that y is described by the (additive noise) regression model,

$$y = f(x_1, \ldots, x_p) + \epsilon \tag{1}$$

over some domain $D \subset \Re^p$, which contains the data. The function $f(\boldsymbol{x})$ reflects the true but unknown relationship between y and $\boldsymbol{x}$. The random additive error variable ϵ, which is assumed to have mean zero and variance σ_ϵ^2, reflects the dependence of y on quantities other than $\boldsymbol{x}$. The goal of a regression modeling methodology is to formulate a function $\hat{f}(\boldsymbol{x})$ that is a reasonable approximation of $f(\boldsymbol{x})$ over the domain D.

Both parametric and nonparametric regression modeling methodologies provide useful methods for developing regression models. If the correct parametric form of $f(\boldsymbol{x})$ is known, then we can use global parametric regression modeling to estimate a finite number of unknown coefficients. Draper and Smith (1966) discuss classical parametric regression modeling and provide extensive discussion of parametric regression modeling techniques.

The most frequently used and well-known form of parametric regression modeling is ordinary least squares regression, which estimates $f(\boldsymbol{x})$ from (1) using

$$\hat{y} = \hat{f}(\boldsymbol{x}) = Hy \tag{2}$$

where H is the well known projection or 'hat' matrix $H = X(X'X)^{-1}X'$. Parametric regression models require less data than nonparametric regression models and their properties are rapidly computed. However, the proper use of parametric regression modeling requires knowledge of the approximate parametric form of the underlying function $f(\boldsymbol{x})$, which can become increasing difficult as the dimension of the predictor variable space p becomes large.

In this dissertation the approach is focused towards nonparametric regression modeling (see, e.g., Eubank, 1988). It is only assumed that $f(\boldsymbol{x})$ belongs to a general collection of functions and the data is used to determine the final model form and its associated coefficients i.e., the form of $f(\boldsymbol{x})$ is not rigidly specified. The most common nonparametric regression methodologies use local parametric (linear smoothing) approximations, or use spline smoothing approximations, to estimate the underlying function $f(\boldsymbol{x})$ (Thisted, 1988).

One difficulty with applying existing nonparametric regression modeling methodologies to problems of dimension greater than two has been called the *curse-of-dimensionality* (Bellman, 1961). The *curse-of-dimensionality* describes the need for an exponential increase in sample size N for a linear increase in p, in order to densely populate higher-dimensional predictor variable spaces. In effect, the *curse-of-dimensionality* limits the practical ap-

plication of some nonparametric regression modeling methodologies to problems of low dimension.

Linear smoothing is a form of nonparametric regression that estimates $f(\boldsymbol{x})$ from (1) with

$$\hat{f}(\boldsymbol{x}) = S\, y \tag{3}$$

where S is an n by n matrix. As in ordinary least squares regression (2), the matrix S depends only on the X matrix. However, in linear smoothing the S matrix can be some nonlinear form of the X matrix (Thisted, 1988). In general, linear smoothers compute the estimate of $f(\boldsymbol{x})$ at $\boldsymbol{x}_i$ using some localized neighborhood of data around $\boldsymbol{x}_i$. Some common linear smoothers include running means, kernel smoothing and running lines (see, e.g., Altman, 1987 and Cleveland, 1979). Although a nonparametric regression model using linear smoothing is rapidly computed, the estimate of $f(\boldsymbol{x})$ can be poor at the extremes of the predictor variable space due to the endpoint behavior of the linear smoother. In addition, the *curse-of-dimensionality* limits the practical application of some linear smoothers to a low-dimensional setting, i.e., p is small. Altman (1987) found that some linear smoothers systematically overestimate (undersmooth) or underestimate (oversmooth) the estimate for $f(\boldsymbol{x})$ when serial correlation is present in the data. Serial correlation in the data can even plague more sophisticated nonlinear smoothers, such as SUPERSMOOTHER (Friedman 1984).

Spline smoothing approximations are a special form of linear smoothing (3) which are particularly attractive as nonparametric regression models because they arise as the solutions to optimization problems closely related to least squares and maximum likelihood (Thisted, 1988). Silverman (1985) views spline smoothing approximations as a span between parametric and nonparametric regression methodologies. An excellent survey and discussion of splines in statistics is provided in papers by Wegman and Wright (1983) and Silverman (1985).

Roughness penalty methods and regression splines are two forms of spline smoothing. Spline smoothing approximations that use roughness penalty methods to estimate $f(\boldsymbol{x})$ are very robust regression modeling methodologies. However, roughness penalty methods are

hampered by the *curse-of-dimensionality* and the large number of coefficients that must be computed for large p. Regression splines seek to overcome the difficulties of roughness penalty methods but still require a methodology for selecting the number and location of the spline knots for the regression model.

Multivariate Adaptive Regression Splines (MARS) (Friedman, 1988) is a new method of flexible nonparametric regression spline modeling that appears to be an improvement over existing regression modeling methodologies when using moderate sample sizes N and predictor spaces with dimension $p > 2$. In the regression context, MARS can be conceptualized as a generalization of a recursive partitioning strategy (Morgan and Sonquist, 1963; Breiman et al., 1984) that uses regression splines in lieu of other simple fitting functions. Given a set of predictor variables, MARS fits a model in the form of an expansion in product spline basis functions of predictors chosen during a forward and backward recursive partitioning strategy. Although MARS is a computationally intensive regression methodology, it provides a systematic (automatic) approach to regression model building that can produce continuous models for high-dimensional data with multiple partitions and predictor variable interactions.

Although MARS is capable of regression modeling in low-dimensional environments, i.e., those for which $p \leq 2$, its primary advantages exist in higher-dimensional predictor spaces where, as discussed above, many regression methodologies are plagued by the *curse-of-dimensionality*. The *curse-of-dimensionality* cannot be overcome if the data used in constructing $f(\boldsymbol{x})$ exhibits no special structure (Friedman, 1988). However, in general, this is not the case. Thus, MARS attempts to overcome the *curse-of-dimensionality* by exploiting the localized low-dimensional structure of the data (where it exists) used in constructing $\hat{f}(\boldsymbol{x})$. Note that in this dissertation the approach taken to explain and apply MARS is geometric in nature; the focus is on the iterative formation of overlapping subregions in the domain D of the predictor variables. Each subregion of the domain is associated with a product spline basis function. MARS approximates the unknown function $f(\boldsymbol{x})$ using the set of product spline basis functions associated with the overlapping subregions of the domain.

What about the use of MARS in a time series setting? By letting the predictor variables in the MARS algorithm for the τth value in the time series $\{X_\tau\}$ be its lagged values, i.e.,

$X_{\tau-1}, X_{\tau-2}, \ldots, X_{\tau-p}$, one obtains an adaptive spline threshold autoregressive (ASTAR) time series model. In the multivariate time series setting, i.e., where the predictor variables are not only the lagged values of the object time series but also the lagged values of other cross-correlated time series, the application of MARS results in a semi-multivariate ASTAR (SMASTAR) time series model. Thus the MARS methodology gives a new method for nonlinear modeling of univariate and multivariate time series and a systematic way of fitting the model to the data. The ASTAR and SMASTAR methodologies extend the threshold autoregression (TAR) methodology developed by Tong (1990) and seem well suited for taking into account the complex interactions among the univariate or multivariate, lagged predictor variables of a time series system.

A significant feature of this application of MARS is its ability to produce nonlinear models with limit cycles when modeling time series data that exhibit periodic behavior. In a physical context, limit cycles represent a stationary state of sustained oscillations, a satisfying behavior for any model of a time series with periodic behavior. Many time series such as the Canadian Lynx data, Wolf sunspot data, and many riverflow data sets exhibit 'periodic' behavior. The Lynx data and Wolf sunspot data are quasi-periodic. However, riverflow data is frequently tied to a fixed yearly oscillation that can dominate the structure of the time series.

B. CONTRIBUTIONS OF THIS DISSERTATION

Much as Yule's (1927) application of least squares regression to linear time series motivated the development of linear autoregressive (AR) modeling, the application of multivariate adaptive regression splines (MARS) to time series systems provides a new and innovative approach for nonlinear time series modeling. The application of MARS to time series systems is a major contribution of this thesis. The systematic (automatic) approach for model building provided by MARS gives an interpretable representation for a nonlinear time series modeling methodology called adaptive spline threshold autoregression (ASTAR) for univariate time series systems and semi-multivariate ASTAR (SMASTAR) for multivariate time series systems. However, the functional form of an ASTAR or SMASTAR model, with the combination of different predictor variables and multiple partitions of the predictor variable space, makes their straightforward interpretation and analysis difficult. In this

regard a graphical and hierarchical representation was developed to permit interpretation and analysis of ASTAR and SMASTAR models.

The ASTAR and SMASTAR methodologies turn out to be a generalization of, and extension of, the nonlinear threshold autoregressive (TAR) methodology developed by Tong (1990). The development of TAR models in the late seventies provided a basis for the 'practical' modeling and investigation of nonlinear univariate and multivariate time series systems (Tong, 1980). Univariate and semi-multivariate TAR models are general enough to capture some non-linear phenomena (such as limit cycles), provide predictive capability, appear to improve upon linear models when used to model nonlinear systems, and provide a much wider class of time series models than available previously. However, in general, TAR models are piecewise, discontinuous, linear autoregressive time series models of disjoint subregions in the domain of the predictor variables. Also, the ability of the TAR methodology to identify the complex interactions between cross-correlated lagged predictor variables, especially in the case of a multivariate time series system, is limited by the threshold selection process. In contrast, ASTAR and SMASTAR models provide a more general class of nonlinear time series models that are continuous in the domain of the predictor variables. The systematic (automatic) approach for developing ASTAR and SMASTAR models seems well suited for taking into account the complex interactions among the univariate and multivariate lagged predictor variables of a time series system. When used for prediction, ASTAR and SMASTAR models are a significant improvement over other existing nonlinear models of the time series investigated in this dissertation.

To facilitate the application of MARS to time series systems, Fortran program subroutines were developed. The input programs permit the user to identify and enter the necessary information for initiating the MARS methodology in a time series setting. The output programs provide ASTAR and SMASTAR model output in a form that facilitates model analysis. In addition, various subroutines written in APL are available to permit graphical and statistical analysis of ASTAR and SMASTAR models using programs such as IBM's GRAFSTAT.

One difficulty that is often faced during regression modeling is the problem of model selection i.e., choosing the appropriate model predictor variables and model dimension. Friedman (1988) uses a modified form of generalized cross validation (GCV^*), first suggested

by Craven and Wahba (1979), for model selection within MARS. However, one question that immediately develops is whether GCV^* is the 'best' criterion for model selection when using serially and cross-correlated time series data. Other model selection criteria, such as Akaike's Information Criterion (AIC) (Akaike, 1974), have been suggested for model development in a standard linear time series setting. Using simulations and mean squared error (MSE) as a performance measure, it is shown that other model selection criterion, in particular the SC (Schwarz, 1978; Rissanen, 1987) criterion, are an improvement over the GCV^* criterion when modeling time series with MARS.

C. OUTLINE OF THIS DISSERTATION

Chapter II provides an introduction to the recursive partitioning and regression spline methodologies that form the foundation for the development of the MARS methodology. This is followed by the development of the ASTAR time series model that results when the MARS algorithm is applied to univariate time series. Simulations are used to demonstrate the ability of ASTAR to detect and model simple linear and nonlinear time series. As an example of ASTAR modeling in a more difficult setting, the last section of Chapter II considers the widely studied yearly Wolf sunspot numbers, a nonlinear time series with periodic behavior. When used for prediction, ASTAR models are a significant improvement over other existing nonlinear models of the Wolf sunspot numbers. Chapter III discusses the semi-multivariate time series extension of ASTAR (called SMASTAR for semi-multivariate adaptive spline threshold autoregression) i.e., the univariate time series to be modeled not only has its own lagged variables as predictors, but also the lagged variables of other related time series. This approach seems well suited for taking into account the complex interactions among multivariate, cross-correlated, lagged predictor variables of a time series system. Analysis of an Icelandic river using past riverflow, temperature and precipitation is used as an example to demonstrate this extension of MARS. The use of this example is predicated on the fact that riverflow time series are very difficult to model because they are frequently nonlinear and nonnormal, and also because this Icelandic riverflow was analyzed by Tong et al. (1985) using semi-multivariate TAR models. Chapter IV explains the development of Fortran subroutines to permit the application of MARS to univariate and semi-multivariate time series systems. An example is provided using 12 years of daily sea-surface temperatures,

a large univariate time series with periodic behavior. Chapter V is a discussion of the problem of model selection within MARS. Using simulations and mean squared error (MSE) as a performance measure, it is shown that other model selection criterion, in particular the *SC* (Schwarz, 1978; Rissanen, 1987) criterion, are an improvement over the GCV^* criterion used in MARS by Friedman when modeling time series.

D. SUMMARY

MARS is a new nonparametric modeling methodology, due to Friedman, that utilizes low-order regression spline modeling and a modified recursive partitioning strategy to exploit the localized low-dimensional behavior of the data used to construct $\hat{f}(\boldsymbol{x})$. MARS is a computationally intensive regression methodology that selects a regression model using exhaustive search procedures. However, MARS provides a systematic (automatic) regression methodology for deriving nonlinear threshold models for high-dimensional data that are naturally continuous in the domain of the predictor variables and can have multiple partitions and predictor variable interactions.

By letting the predictor variables in MARS be lagged values of a univariate time series, one obtains an adaptive spline threshold autoregressive (ASTAR) time series model, which is a new, computationally intensive method for the systematic nonlinear modeling of a univariate time series system. The MARS methodology is easily extended to the semi-multivariate nonlinear modeling of a single object time series in a multivariate times series system (SMASTAR); this approach is well suited to take into account the complex and possibly nonlinear interactions among cross-correlated, lagged predictor variables of a multivariate time series system. Also, simulations suggest other model selection criterion, such as the *SC* (Schwarz, 1978; Rissanen, 1987) criterion, for use within MARS when modeling in a time series setting. Fortran programs are available for implementing MARS in a time series setting; the drivers for the Fortran programs are given in Appendices A, B, and C.

II. NONLINEAR MODELING OF UNIVARIATE TIME SERIES USING MULTIVARIATE ADAPTIVE REGRESSION SPLINES (MARS)

A. INTRODUCTION

This chapter introduces MARS, due to Friedman (1988), a new methodology for regression analysis which, when applied to nonlinear time series, extends the nonlinear threshold autoregression methodology (TAR) developed by Tong (1985). To motivate the development of the MARS procedure, Sections B and C of this chapter briefly review recursive partitioning and regression splines. Section D of this chapter is a discussion of Friedman's innovations used to develop MARS. An algorithm for implementing MARS is addressed in Section E of this chapter. The application of MARS to univariate time series for developing adaptive spline threshold autoregression (ASTAR) models is discussed in Section F of this chapter. The final part of Section F is an application of MARS to the Wolf sunspot numbers, an often studied univariate time series with periodic behavior.

The approach taken to explain and apply MARS is geometric in nature, i.e., the iterative formation of overlapping subregions in the domain D of the predictor variables. Each one of the domain's subregions, developed using a modification of a forward and backward recursive partitioning strategy, is associated with a product spline basis function. MARS approximates the unknown function $f(\boldsymbol{x})$ (in equation 1) using the set of product spline basis functions associated with the overlapping subregions of the domain.

A significant feature of ASTAR when modeling time series data with periodic behavior is its ability to produce continuous models for the regression function with underlying sustained oscillations (limit cycles). An initial analysis of the yearly Wolf sunspot numbers using ASTAR produced several models with underlying limit cycles. When used for prediction, ASTAR models are a significant improvement over other existing nonlinear models of the Wolf sunspot numbers.

B. RECURSIVE PARTITIONING (RP)

The origin of recursive partitioning regression modeling methodology appears to date to the development and use of the AID (Automatic Interaction Detection) program by Morgan and Sonquist in the early 1960's. More recent extensions and contributions were made by Breiman et al. (1984). In Subsection 1 recursive partitioning (RP) is explained using recursive splitting of established subregions. In Subsection 2 recursive partitioning is then recast equivalently as an expansion in a set of basis functions. The latter explanation of recursive partitioning may be considered a precursor to MARS.

1. RP: Recursive Splitting of Established Subregions

Let the response variable y depend in some unknown way on a vector of p predictor variables $\boldsymbol{x} = (x_1, \ldots, x_p)$, that is modeled with (1). Assume there are N samples of y and $\boldsymbol{x}$, namely $\{y_i, \boldsymbol{x}_i\}_{i=1}^N$. Let $\{R_j\}_{j=1}^S$ be a set of S disjoint subregions of $D \subset \Re^p$ such that $D = \bigcup_{j=1}^{S} R_j$. Given the subregions $\{R_j\}_{j=1}^S$, recursive partitioning estimates the unknown function $f(\boldsymbol{x})$ at $\boldsymbol{x}$ with

$$\hat{f}(\boldsymbol{x}) = \hat{f}_j(\boldsymbol{x}) \quad \text{for} \quad \boldsymbol{x} \in R_j, \tag{4}$$

where the function $\hat{f}_j(\boldsymbol{x})$ estimates the true but unknown function $f(\boldsymbol{x})$ over the R_jth subregion of D. In recursive partitioning, $\hat{f}_j(\boldsymbol{x})$ is frequently taken to be the constant function (Morgan and Sonquist, 1963 and Breiman et al., 1984) although linear functions have been proposed without much success (Breiman and Meisel, 1976). For the purpose of explaining MARS, $\hat{f}_j(\boldsymbol{x})$ is taken to be a constant function,

$$\hat{f}_j(\boldsymbol{x}) \;=\; c_j \;\; \forall \;\; \boldsymbol{x} \in R_j, \tag{5}$$

where each c_j is chosen to minimize the jth component of the residual-squared-error (badness-of-fit),

$$BOF[\hat{f}_j(\boldsymbol{x})] = \min_{c_j} \sum_{\boldsymbol{x}_i \in R_j} (y_i - c_j)^2. \tag{6}$$

Since the subregions of the domain D are disjoint, each c_j will be the sample mean of the y_i's whose $\{\boldsymbol{x}_i\}_{i=1}^N \in R_j$.

In general, the recursive partitioning model is the result of a 2-step procedure that starts with the single subregion $R_1 = D$. The first, or forward, step of the algorithm uses recursive splitting of established subregions to iteratively produce a large number of disjoint subregions $\{R_j\}_{j=2}^{M}$, for $M \geq S$, where M is chosen by the user. The second, or backward, step of the algorithm reverses the first step and trims an excess $(M - S)$ subregions from the model using a criterion that evaluates both the model fit and the number of subregions in the model. The goal of the 2-step procedure is to use the data to select a good set of subregions $\{R_j\}_{j=1}^{S}$ together with the constant functions c_j that estimate $f(\boldsymbol{x})$ over each subregion of the domain.

To facilitate understanding of the recursive partitioning algorithm we examine the forward-step procedure for an example problem using $p = 3$ predictor variables, and $M = 5$, the maximum number of forward-step subregions. Let $v = 1, \ldots, p$ index the predictor variables and $k = 1, \ldots, n$, index the ordered sample values of a predictor variable x_v in subregion R_j. For our purposes we use $BOF_m = \sum_{j=1}^{m} BOF[\hat{f}_j(\boldsymbol{x})]$ as the forward-step measure of fit for a recursive partitioning model with m subregions, and we restrict the set of candidate partition points to the actual sample values, $x_{v,k}$. Note that $x_{v,k}$ represents the kth serially-ordered sample value of the vth predictor variable, while x_v alone denotes the running values of the vth predictor variable. At the start of the forward-step recursive partitioning algorithm, R_1 is the entire domain D and the single subregion estimate for $f(\boldsymbol{x})$ is

$$\hat{f}(\boldsymbol{x}) = \hat{f}_1(\boldsymbol{x}) = c_1 = \frac{1}{N} \sum_{i=1}^{N} y_i. \tag{7}$$

The forward-step measure of fit for the single subregion recursive partitioning model is

$$BOF_1 = \sum_{i=1}^{N} (y_i - c_1)^2. \tag{8}$$

The initial recursion, $m = 2$, for the forward-step algorithm selects a partition point t^* that best splits subregion R_1 into two disjoint sibling subregions. The method for discovering t^* is a straightforward exhaustive search; evaluate *every* sample value $x_{v,k}$ (for $v = 1, \ldots, p; k = 1, \ldots, n$) as a candidate partition point to determine which one minimizes the remaining badness-of-fit for a $m = 2$ subregion model. For example, let $t = x_{1,15}$ identify

a candidate partition point for predictor variable x_1. The area in parent subregion R_1 to the left of t, $x_1 < t$, resides in proposed sibling subregion $R_{1,l}$. The area to the right of t, $t \leq x_1$, resides in proposed sibling subregion $R_{1,r}$. Given the proposed split of R_1 along $t = x_{1,15}$, we evaluate the model using BOF_m for a $m = 2$ subregion model, i.e.,

$$BOF_2 = \min_{c_l} \sum_{\boldsymbol{x}_i \in R_{1,l}} (y_i - c_l)^2 \; + \; \min_{c_r} \sum_{\boldsymbol{x}_i \in R_{1,r}} (y_i - c_r)^2. \tag{9}$$

Using the indices v and k, the exhaustive search sequentially evaluates all possible partition points for each predictor variable in R_1 (which here is equal to D).

For our example problem, let the partition point $t^* = x_{2,25}$ identify the split of subregion R_1 that minimizes the forward-step fit criterion BOF_m for a $m = 2$ subregion recursive partitioning model. We use $x_{2,25}$ to create two new disjoint subregions during the split and elimination of the old parent region, which we now call $R_{1\bullet}$. First, the area in parent subregion $R_{1\bullet}$ to the left of t^* (i.e., $x_2 < t^*$) is assigned to sibling subregion R_2, while the area to the right of t^* (i.e., $t^* \leq x_2$) is reconstituted as subregion R_1. The creation of the two new disjoint subregions R_1 and R_2, and the elimination of the old parent subregion $R_{1\bullet}$, increase by one the number of disjoint subregions that partition D completing the initial recursion of the forward-step procedure. Thus, the two-subregion recursive partitioning estimate of $f(\boldsymbol{x})$ for our example problem is

$$\hat{f}(\boldsymbol{x}) = c_j \;\text{ if }\; \boldsymbol{x} \in R_j, \;\text{ for } j = 1,2 \tag{10}$$

where (since we are splitting the domain D on only x_2's dimension),

$$\boldsymbol{x} \in \begin{cases} R_1 & \text{if} \quad x_2 \geq x_{2,25} \\ R_2 & \text{if} \quad x_2 < x_{2,25}. \end{cases}$$

Note that the form of the recursive partitioning model (4) did not change during the recursion, but only the number of disjoint subregions that partition D.

The recursions $m = 3, \ldots, M = 5$ of the forward-step algorithm repeat the first recursion with one exception. The exhaustive search is now conducted to identify the best split (minimizing BOF_m) for *one and only one* of the subregions from the current $m - 1$ subregion model. Each recursion's partition point t^* is selected as before, after

an evaluation of all potential partition points for each predictor variable in the existing subregions $\{R_j\}_{j=1}^{m-1}$ of the model. The recursive splitting continues until the domain D is partitioned into $M = 5$ disjoint subregions $\{R_j\}_{j=1}^{5}$. Upon completion of the forward-step recursive partitioning algorithm, a backward-step algorithm trims excess subregions using a criterion that evaluates both fit and the number of subregions in the model. (See Friedman (1988) for a discussion of the backward-step algorithm). Completion of the backward-step procedure results in the final recursive partitioning model with $\{R_j\}_{j=1}^{S}$ subregions $(S \leq M)$.

2. RP: An Expansion in a Set of Basis Functions

While the initial approach to understanding recursive partitioning is through recursive splitting, it is recast now in an equivalent form to provide a reference for explaining the MARS methodology. The central idea is to formulate the recursive partitioning model as an additive model of functions from disjoint subregions. Also, we associate the operation of subregion-splitting with the operation of step-function multiplying. The new approach approximates the unknown function $f(\boldsymbol{x})$ at $\boldsymbol{x}$ with an expansion in a set of basis functions from disjoint subregions $\{R_j\}_{j=1}^{S}$:

$$\hat{f}(\boldsymbol{x}) = \sum_{j=1}^{S} c_j B_j(\boldsymbol{x}), \tag{11}$$

where

$$B_j(\boldsymbol{x}) = I\left[\boldsymbol{x} \in R_j\right],$$

and $I[\cdot]$ is an indicator function with value 1 if its argument is true and 0 otherwise. The constant function c_j estimates the true, but unknown function, $f(\boldsymbol{x})$ over the R_jth subregion of D, and $B_j(\boldsymbol{x})$ is a basis function that indicates membership in the R_jth subregion of D. We call $B_j(\boldsymbol{x})$ a *basis function* because it restricts contributions for $\hat{f}(\boldsymbol{x})$ to those values of $\boldsymbol{x}$ in the R_jth subregion of D. The approximation of the unknown function $f(\boldsymbol{x})$ at $\boldsymbol{x}$ in (4) and (11) are equivalent: the subregions $\{R_j\}_{j=1}^{S}$ are the same disjoint subregions of the domain D, and the constant functions $\{c_j\}_{j=1}^{S}$ are the same constant functions that estimate $f(\boldsymbol{x})$ over each subregion.

During each search for a partition of a subregion R_j using an expansion in a set of basis functions (11), the selection of a candidate partition point creates a particular functional form for $\hat{f}(\boldsymbol{x})$ that we call g in the following algorithm. Let

$$H[\eta] = \begin{cases} 1 & \text{if } \eta \geq 0 \\ 0 & \text{otherwise,} \end{cases} \tag{12}$$

be a step function (which returns a value of 1 if η is not negative, and 0 otherwise). Following Friedman (1988), an algorithm to implement the forward-step recursive partitioning procedure using an expansion in a set of basis functions is:

Recursive Partitioning Algorithm (Forward-Step) (13)

$R_1 = D,\ B_1(\boldsymbol{x}) = 1$ (a)

For each subregion R_m, $m = 2$ to M do: (b)

  bof$^* = \infty,\ j^* = 0,\ v^* = 0,\ t^* = 0$ (c)

  For each established subregion R_j, $j = 1$ to $m-1$ do: (d)

    For each predictor variable x_v in R_j, $v = 1$ to p do: (e)

      For each data value $x_{v,k}$ in R_j, $t = x_{v,k=1}$ to $x_{v,k=n}$ do: (f)

        $g = \left(\sum_{d \neq j} c_d B_d(\boldsymbol{x})\right) + c_m B_j(\boldsymbol{x}) H[t - x_v] + c_j B_j(\boldsymbol{x}) H[x_v - t]$ (g)

        bof = BOF_m (h)

        if bof < bof* then bof* = bof ; $j^* = j$; $v^* = v$; $t^* = t$ end if (i)

      end for

    end for

  end for

  $R_m \leftarrow \{R_{j^*} : (t^* - x_{v^*}) > 0\}$ (j)

  $R_{j^*} \leftarrow \{R_{j^*} : (x_{v^*} - t^*) \geq 0\}$ (k)

end for

end

The forward-step recursive partitioning algorithm is initialized with the first subregion R_1 equal to the entire domain D (13a). The outer loop (13b) controls the iterative creation of the subregions $\{R_m\}_{m=2}^{M}$. Next, the dummy variables (13c) for the evaluation of

the fit procedure bof*, region j^*, predictor variable v^*, and partition point t^* are initialized in preparation for identifying the next partition of an established subregion $\{R_j\}_{j=1}^{m-1}$. The three nested inner loops (13d-13f) perform the exhaustive search for the next partition point by iteratively searching across all established subregions (13d), all predictor variables (13e), and all values of the predictor variables in the jth subregion (13f). Given the investigation of a partition point t for a predictor variable x_v in subregion R_j, the function g (13g), with parameter vector $\boldsymbol{c} = (c_1, \ldots, c_m)$, is the current candidate for a recursive partitioning model estimate of $f(\boldsymbol{x})$ in the mth iteration of the forward-step procedure. The first term in (13g) includes all subregions except subregion R_j. The last two terms in (13g),

$$c_m B_j(\boldsymbol{x}) H[t - x_v] + c_j B_j(\boldsymbol{x}) H[x_v - t],$$

reflect the proposal to divide the parent subregion R_j into two disjoint sibling subregions using the step functions $H[t - x_v]$ and $H[x_v - t]$ to identify each $\boldsymbol{x}$'s location with respect to the partition point t. Next, BOF_m (13h) is the forward-step measure of fit that evaluates the function g with respect to the data. Information for the best yet discovered partition, predictor variable, and subregion is retained (step 13i) as the search continues for the best partition of an established subregion $\{R_j\}_{j=1}^{m-1}$ in the mth iteration. Completion of the mth iteration's search results in the division (and elimination) of the old parent subregion R_{j^*} into two disjoint sibling subregions (13j and 13k) based on x_{v^*}'s location with respect to the partition point t^*. The iterations continue until the domain D is partitioned into M disjoint subregions $\{R_j\}_{j=1}^{M}$.

Each basis function $B_j(\boldsymbol{x})$ identifies membership in the R_jth subregion of D and is the result of the product of step functions whose partition points define the subregion R_j. For example, let $D \in \Re^2$ and R_5 be a subregion formed from the sequence of step functions $H[x_1 - t_1^*], H[t_2^* - x_2], H[x_2 - t_3^*]$ and $H[t_4^* - x_1]$ where $\{t_i^*\}_{i=1}^4$ is 0,1,0,1 respectively. Then the basis function $B_5(\boldsymbol{x})$ is,

$$B_5(\boldsymbol{x}) = H[x_1 - 0] \times H[1 - x_2] \times H[x_2 - 0] \times H[1 - x_1], \tag{14}$$

which delineates the subregion R_5 as a unit square in $\Re^2$. The basis function $B_5(\boldsymbol{x}) = 1$ if $0 \leq x_1 \leq 1$ and $0 \leq x_2 \leq 1$, and 0 otherwise.

In recursive partitioning, the subregions $\{R_j\}_{j=1}^{S}$ *are disjoint.* Each data point $\boldsymbol{x}$ belongs to only one subregion R_j. Therefore, the estimate of $f(\boldsymbol{x})$ over each subregion R_j is restricted to the functional form for $\hat{f}_j(\boldsymbol{x})$, which in this discussion is the constant function c_j. However, as we will address in Section D, MARS *has overlapping subregions.* The estimate of $f(\boldsymbol{x})$ over subregion R_j may be obtained as a sum of multiple functional forms.

Recursive partitioning is a very powerful regression modeling methodology that is rapidly computed, especially if $\hat{f}_j(\boldsymbol{x})$ is a constant function c_j. Each forward step of the algorithm (13) partitions *one and only one* subregion of the domain on an influential variable x_{v^*}. This procedure increasingly localizes the activity of the predictor variables with respect to the response variable y. However, there are several drawbacks to using recursive partitioning as a regression modeling technique:

- Recursive partitioning models have disjoint subregions giving rise to discontinuities at subregion boundaries. This is disconcerting if we believe $f(\boldsymbol{x})$ is continuous.
- Recursive partitioning has an innate *inability* to adequately estimate functions $f(\boldsymbol{x})$ that are linear or additive. This is due to the recursive division of established subregions during the forward-step procedure that automatically produces predictor variable interactions, i.e., terms of the form cx_ix_j, unless all successive partitions occur on the same predictor variable.
- The form of the recursive partitioning model (11), which is an additive combination of functions of predictor variables in disjoint regions, makes estimation of the true form of the unknown function $f(\boldsymbol{x})$ difficult for large p.

C. REGRESSION SPLINES

The development of a regression spline model offers another method for explaining MARS. Silverman (1985) views spline functions as an attractive approach to modeling that may be thought of as a span between parametric and nonparametric regression methodology. For simplicity, define a qth-order polynomial function in the unknown $x \in D \subset \Re^1$ with coefficients c_l as follows:

$$p_q(x) = \sum_{l=0}^{q} c_l x^l \quad \text{for } x \in D. \tag{15}$$

The polynomials in (15) are smooth and easy to manipulate. However, global fitting of data with a polynomial model may require higher-order terms having unacceptable fluctuations.

This observation leads us to divide the domain D into smaller subregions R_j to permit the use of (different) polynomial functions of relatively low order within each subregion.

Let $[a,b] = D \subset \Re^1$, and let $\Delta_S = \{t_1, \ldots, t_{S-1}\}$ denote an ordered partition of $[a,b]$ into S disjoint subregions $a = t_0 < t_1 < \ldots < t_{S-1} < t_S = b$. Denote each disjoint subregion by $R_j = [t_{j-1}, t_j]$, for $j = 1, \ldots, S$. Let $C^q[D]$ represent the set of all continuous functions in D whose $q-1$ derivatives are also continuous. Using j as a subscript to index the subregions, we define a spline function $s^q_{\Delta_S}$ as a set of S piecewise qth-order polynomial functions whose function values and first $q-1$ derivatives agree at their partition points, i.e.,

$$s^q_{\Delta_S}(x) = \sum_{j=1}^{S} p_{q,j}(x)\, I[x \in R_j], \tag{16}$$

with the restriction that $s^q_{\Delta_S}(x) \in C^q[D]$.

There are several approaches for implementing (Wegman and Wright, 1983) splines within a regression setting. One approach is to write the regression model (1) as a piecewise regression spline model,

$$y = s^q_{\Delta_S}(x) + \epsilon, \tag{17}$$

where ϵ is assumed to have mean zero, variance σ^2_ϵ and to be independent of $s^q_{\Delta_S}(x)$. Moreover, $s^q_{\Delta_S}(x)$ estimates $f(x)$ according to (16).

Given a set of partitions points Δ_S, Smith (1979) has shown that a different and more useful regression spline model may be written using plus (+) functions. The **plus function** is defined as

$$u_+ = \begin{cases} u & \text{if } u > 0 \\ 0 & \text{if } u \leq 0. \end{cases} \tag{18}$$

Again, let $[a,b] = D \subset \Re^1$. However, we now let $\Delta_{So} = \{t_1, \ldots, t_{S-1}\}$ define an *ordered* partition of $[a,b]$ into S *overlapping subregions* and denote the S *overlapping subregions* as $R_j = [t_{j-1}, t_S]$, for $j = 1, \ldots, S$. Let l index the order of the polynomial terms in each subregion of the domain and c_{jl} denote the coefficient for the lth term of the polynomial

function in the $(j+1)$st subregion of a spline model. The use of plus functions results in a truncated regression spline model functionally equivalent to the piecewise regression spline model (16) as follows:

$$y = \sum_{l=0}^{q} c_{0l} x^l + \sum_{j=1}^{S-1} c_{jq}[(x - t_j)_+]^q \; + \; \epsilon, \quad q \geq 1, \tag{19}$$

where ϵ is assumed to have mean zero, variance σ_ϵ^2 and is independent of $s^q_{\Delta_S}(x)$, and q is assumed to be greater than or equal to one. Since the partition points of the set Δ_{S_o} are ordered, the number of overlapping truncated spline functions with nonzero values increases by one as we move to the right, across each partition point t_j. Figure 1 compares the different forms for $(q = 2)$-order piecewise (16) [top] and truncated (19) [bottom] spline functions, both with a single partition point at x= 1, that equivalently define a line y = $f(\boldsymbol{x})$. In the top plot the line y from $0 \leq x \leq 2$ is defined by two disjoint 2nd-order polynomial functions that are shown using different triangular symbols; one 2nd-order polynomial function shown as $\bigtriangledown\,\bigtriangledown\cdots$ in subregion [0,1) and one 2nd-order polynomial function shown as $\bigtriangleup\,\bigtriangleup\cdots$ in subregion [1,2]. In the bottom plot the line y in the first subregion [0,1) is also defined by a single 2nd-order polynomial function shown as $\bigtriangledown\,\bigtriangledown\cdots$. However, in the second subregion [1,2] the line y is defined as the sum of two overlapping 2nd-order polynomial functions; the first a 2nd-order polynomial function overlapping from the first subregion shown as $\bigtriangledown\,\bigtriangledown\cdots$ and the second a truncated 2nd-order polynomial function shown as $\bigtriangleup\,\bigtriangleup\cdots$. Both the piecewise (16) and truncated (19) spline functions equivalently define the line y.

The key point of this section is that once the number and the values of the partition points $\{t_j\}_{j=1}^{S-1}$ *are fixed, the qth-order truncated regression spline model (19) with those partition points is a linear model whose coefficients* **c** *are determined by straightforward least-squares regression. Nevertheless, the major difficulty in implementing a qth-order regression spline model is in choosing the number and values of the partition points.*

We have defined regression spline models in $\Re^1$. The extension to higher dimensions for $p > 1$ predictor variables is usually accomplished through products of univariate spline functions. Nevertheless, regression using products of univariate spline functions suffers from the *curse-of-dimensionality* discussed previously. From the perspective of regression splines, MARS attempts to overcome the *curse-of-dimensionality* by using a modified recursive

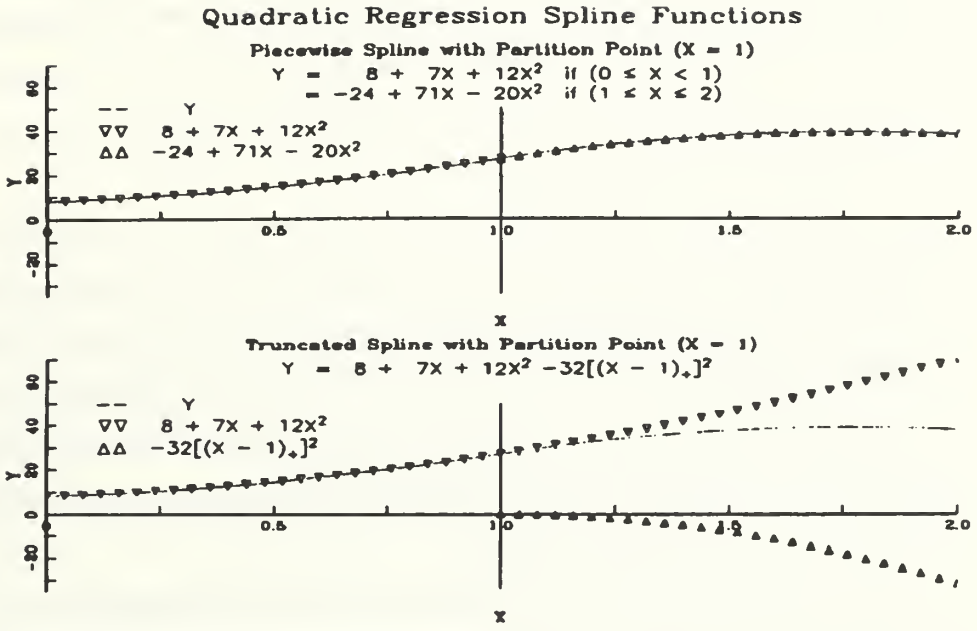

Figure 1. The different forms for a piecewise (16) and truncated (19) spline function using ($q = 2$)-order splines over the region $D = [0, 2]$ with a single partition point at x= 1. In the top plot, the line y is defined by two different disjoint 2nd-order polynomial functions; one 2nd-order polynomial function for the region [0,1) shown as $\nabla\,\nabla\cdots$ and one 2nd-order polynomial function for the region [1,2] shown as $\Delta\,\Delta\cdots$. In the bottom plot the line y in the region from [0,1) is again defined by the single 2nd-order polynomial function $8 + 7x+12x^2$ shown as $\nabla\,\nabla\cdots$. However, in the subregion from [1,2], the line y is defined as the sum of two overlapping 2nd order polynomial functions; $8 + 7x+12x^2$ that continues from the first subregion and the truncated 2nd-order polynomial function $-32[(x-1)_+]^2$ shown as $\Delta\,\Delta\cdots$.

partitioning strategy to select partitions of the domain. This permits MARS to exploit the localized, low-dimensional structure of the data using linear (i.e., $q = 1$ order) truncated, multidimensional regression spline functions.

D. FRIEDMAN'S INNOVATIONS FOR RECURSIVE PARTITIONING

Recursive partitioning and regression splines have tremendous power for modeling in high-dimensional environments. Each approach also presents difficulties when applied; recursive partitioning has discontinuities, variable interactions and poor model interpretation, and regression splines battle the *curse-of-dimensionality* and lack a methodology to optimally select its many parameters.

Two aspects of the recursive partitioning algorithm (13) contribute to the difficulties of its application in a high-dimensional setting. The iterative division and elimination of the parent region when creating its sibling subregions causes difficulty in estimating linear and additive functions. The discontinuous nature of the step function $H[\eta]$ (12) when applied in each linear regression of the forward-step recursive partitioning algorithm (13g) causes the lack of continuity. Together, these characteristics make interpretation of the recursive partitioning model difficult at best.

To overcome recursive partitioning's difficulty in estimating linear and additive functions, Friedman (1988) proposes that the parent region is not eliminated (as in recursive partitioning) during the creation of its sibling subregions. Thus, in future iterations both the parent and its sibling subregions are eligible for further partitioning. An immediate result of retaining parent regions is overlapping subregions of the domain. Also, each parent region may have multiple sets of sibling subregions. With this modification, recursive partitioning can produce linear models with the repetitive partitioning of the initial region R_1 by different predictor variables. Additive models with functions of more than one predictor variable can result from successive partitioning using different predictor variables. This modification also allows for multiple partitions of the same predictor variable from the same parent region.

The above modified recursive partitioning algorithm in which the parent region is maintained results in a class of models with greater flexibility than permitted in recursive partitioning. However, the modified approach is still burdened with the discontinuities

caused by the step function $H[\eta]$. To alleviate this difficulty, Friedman proposes to replace the step function $H[\eta]$ in the model formulation step (13g) with linear (i.e., $q = 1$ order) regression splines in the form of left (−) and right (+) truncated splines. Let $\boldsymbol{r}_m$ represent a 2-tuple associated with the R_mth subregion whose components identify the direction (left or right), specific predictor variable, and partition point used to create subregion R_m from its parent region. Left and right truncated splines for creating the R_mth and R_{m+1}st subregion from the parent region R_j with a partition point in the domain of x_v at t are defined as

$$\begin{aligned} T_{j,\boldsymbol{r}_m}(\boldsymbol{x}) &= [(t - x_v)_+]^{q=1} = (t - x_v)_+ \quad \text{and} \\ T_{j,\boldsymbol{r}_{m+1}}(\boldsymbol{x}) &= [(x_v - t)_+]^{q=1} = (x_v - t)_+, \end{aligned} \tag{20}$$

where $\boldsymbol{r}_m = (-v, t)$ and $\boldsymbol{r}_{m+1} = (+v, t)$ and $m > j$. The additional subscripts j and m, or j and $m+1$, provide a necessary audit trail for products of truncated splines when interactions are allowed among multiple predictor variables. *Note that the truncated spline functions act in only one dimension although their argument is a vector of predictor variables.*

A modeling approach using linear truncated splines (20) creates a continuous approximating function $\hat{f}(\boldsymbol{x})$ with discontinuities in the first partial derivative of $\hat{f}(\boldsymbol{x})$ at the partition points of each predictor variable in the model. The argument for using *linear* truncated splines (20) is that there is little to be gained in flexibility, and much to lose in computational speed by imposing continuity beyond the function $\hat{f}(\boldsymbol{x})$. Linear truncated splines allow rapid updating of the regression model and its coefficients during each exhaustive search for the next partition of an established subregion. The placement of additional partitions may be used to compensate for the loss of flexibility in using linear truncated splines to estimate $f(\boldsymbol{x})$ over a subregion of the domain.

Implementation of the modifications proposed above to the recursive partitioning algorithm avoids its identified difficulties and results in the MARS algorithm. The MARS algorithm produces a linear ($q = 1$) truncated spline model (19) with overlapping subregions $\{R_j\}_{j=1}^{S}$ of the domain D. Each overlapping subregion of a MARS model is defined by the partition points of the predictor variables from an ordered sequence of linear truncated splines.

Define the product basis function $K_m(\boldsymbol{x})$ as the ordered sequence of truncated splines associated with subregion R_m. The first term of every product basis function is $T_{0,\boldsymbol{r}_1}(\boldsymbol{x}) = 1$, the initialization function associated with R_1. Each additional truncated spline represents the iterative partitioning of a parent region into a sibling subregion. For example, assume the sequence of ordered truncated splines for the parent region R_7 is (1,3,7), which is split using $T_{7,\boldsymbol{r}_m}(\boldsymbol{x})$ to create subregion R_m. The product basis function $K_m(\boldsymbol{x})$ associated with the R_mth subregion for this example is

$$K_m(\boldsymbol{x}) = \underbrace{T_{0,\boldsymbol{r}_1}(\boldsymbol{x}) \times T_{1,\boldsymbol{r}_3}(\boldsymbol{x}) \times T_{3,\boldsymbol{r}_7}(\boldsymbol{x})}_{K_7(\boldsymbol{x})} \times T_{7,\boldsymbol{r}_m}(\boldsymbol{x}). \tag{21}$$

where $m > 7$.

To evaluate $K_m(\boldsymbol{x})$ at $\boldsymbol{x}$ requires the evaluation of each truncated spline in the product basis function at $\boldsymbol{x}$. If any of the truncated spline evaluations at $\boldsymbol{x}$ are zero, then $K_m(\boldsymbol{x})$ at $\boldsymbol{x}$ is 0. Otherwise, the evaluation of $K_m(\boldsymbol{x})$ at $\boldsymbol{x}$ is the product of the truncated splines at $\boldsymbol{x}$. For example, let the ordered truncated splines for $R_5 \in \Re^3$ be (1, 2 and 5) with $\boldsymbol{r}_2 = (2,3)$ and $\boldsymbol{r}_5 = (-3,1)$. The product basis function associated with R_5 is

$$\begin{aligned} K_5(\boldsymbol{x}) = \;& T_{0,\boldsymbol{r}_1}(\boldsymbol{x}) \times T_{1,\boldsymbol{r}_2}(\boldsymbol{x}) \times T_{2,\boldsymbol{r}_5}(\boldsymbol{x}) \\ = \;& 1 \times (x_2 - 3)_+ \times (1 - x_3)_+ \quad = \begin{cases} (x_2 - 3)(1 - x_3) & \text{if } x_2 > 3 \text{ and } x_3 < 1 \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

If $\boldsymbol{x} = \{5,4,.5\} \in R_5$ then $K_5(\boldsymbol{x}) = .5$ and if $\boldsymbol{x} = \{4,3.5,6\} \notin R_5$, then $K_5(\boldsymbol{x}) = 0$.

The level of interaction of the predictor variables associated with R_j is the number of truncated splines (without $T_{0,\boldsymbol{r}_1}(\boldsymbol{x})$) in a product basis function $K_j(\boldsymbol{x})$. A one term product basis function represents a truncated linear relationship of its predictor variable while a two term product basis function represents a truncated 2-way interaction and so on. The number and level of interactions in a MARS model are only limited by the data and the maximum level of interaction (an input parameter) permitted in the MARS algorithm.

The MARS estimate of the unknown function $f(\boldsymbol{x})$ is

$$\hat{f}(\boldsymbol{x}) = \sum_{j=1}^{S} c_j K_j(\boldsymbol{x}), \tag{22}$$

where $\hat{f}(\boldsymbol{x})$ is an additive function of the product basis functions $\{K_j(\boldsymbol{x})\}_{j=1}^{S}$ associated with the subregions $\{R_j\}_{j=1}^{S}$. Since for a given set of product basis functions the values of the partition points, which of course are parameters of the model, are *fixed*, the MARS model (22) is a linear model whose coefficients $\{c_j\}_{j=1}^{S}$ may be determined by straightforward least squares regression.

As in recursive partitioning the objective of the *forward-step* MARS algorithm is to iteratively adjust the vector of coefficient values to best fit the data while identifying the subregions $\{R_j\}_{j=1}^{M}$, for $M \geq S$, whose product basis functions approximate $f(\boldsymbol{x})$ based on data at hand. And again, as in the recursive partitioning procedure, it makes sense to follow the forward step procedure with a backward-step trimming procedure to remove the excess $(M - S)$ subregions from the model whose product basis functions no longer sufficiently contribute to the accuracy of the model fit.

MARS uses residual-squared-error, because of its attractive computational properties, in the forward and backward steps of the algorithm to evaluate model fit and compare partition points. The actual backward fit criterion that is used for final model selection is a modified form of the generalized cross validation criterion (GCV) first proposed by Craven and Wahba (1979). The modified generalized cross validation criterion (GCV^*) used in a MARS model with subregions $\{R_j\}_{j=1}^{M}$ is,

$$GCV^*(M) = \frac{\frac{1}{N}\sum_{i=1}^{N}[y_i - \hat{f}_M(\boldsymbol{x}_i)]^2}{[1 - \frac{C(M)^*}{N}]^2}. \tag{23}$$

The numerator in GCV^* is the average residual-squared-error and the denominator is a penalty term that reflects model complexity. The difference in GCV^* and GCV is in the computation of $C(M)^*$, a model complexity penalty function that is increasing in M. In MARS this modification is necessary to account for the heavy use of the data in determining both the partition points and the coefficients of a final model. The use of other criteria, perhaps more suitable to time series applications, is examined in Chapter V.

E. FORWARD STEP MARS ALGORITHM

The MARS forward-step algorithm (24) results from applying the modifications addressed in Section D to the forward-step recursive partitioning algorithm (13). Again we

initialize $R_1 = D$. However, in MARS we create two new subregions R_m and R_{m+1} and maintain the parent region R_{j^*} during each partition. Also, MARS restricts each sequence of truncated splines from having more than one partition per predictor variable because this creates a nonlinear spline function i.e., one with $q > 1$. MARS enforces this restriction, during the search for the next best partition of a subregion R_j, by excluding from consideration for a partition point any predictor variable already included in the product basis function $K_j(\boldsymbol{x})$. The most notable difference between the RP and MARS algorithms occurs in forming the MARS model. Again following Friedman (1988), the product basis functions $\{K_j(\boldsymbol{x})\}_{j=1}^{m}$ given at (21) and the truncated splines $T_{j,\boldsymbol{r}_m}(\boldsymbol{x})$ and $T_{j,\boldsymbol{r}_{m+1}}(\boldsymbol{x})$ given at (20) replace the basis functions $\{B_j(\boldsymbol{x})\}_{j=1}^{m}$ and the step functions $H[t - x_v]$ and $H[x_v - t]$ from equation (12) in the forward step recursive partitioning algorithm (13g) respectively.

MARS Forward Step Algorithm (24)

$R_1 = D,\ T_{0,\boldsymbol{r}_1}(\boldsymbol{x}) = 1$ (a)

For each subregion R_m, $m = 2$ to M do: (b)

  bof$^* = \infty,\ j^* = 0,\ v^* = 0,\ t^* = 0$ (c)

  For each established subregion R_j, $j = 1$ to $m - 1$ do: (d)

    For each predictor variable x_v in R_j, $v = 1$ to p such that $v \notin K_j(\boldsymbol{x})$ do: (e)

      For each data value $x_{v,k}$ in R_j, $t = x_{v,k=1}$ to $x_{v,k=n}$ do: (f)

        $g = (\sum_d c_d K_d(\boldsymbol{x})) + c_m K_j(\boldsymbol{x}) T_{j,\boldsymbol{r}_m}(\boldsymbol{x}) + c_{m+1} K_j(\boldsymbol{x}) T_{j,\boldsymbol{r}_{m+1}}(\boldsymbol{x})$ (g)

        bof= BOF_m (h)

        if bof $<$ bof* then bof* = bof ; $j^* = j$; $v^* = v$; $t^* = t$ end if (i)

      end for

    end for

  end for

  $R_m \leftarrow \{R_{j^*} : (t^* - x_{v^*}) > 0\}$ (j)

  $R_{m+1} \leftarrow \{R_{j^*} : (x_{v^*} - t^*) \geq 0\}$ (k)

  $m \leftarrow m + 2$ (l)

end for

end

To characterize this MARS forward-step procedure we use the example discussed in Subsection 1 of Section B with $p = 3$ predictor variables, and $M = 5$, the maximum number of forward-step partitions. The MARS algorithm parallels the recursive partitioning algorithm except for the modifications discussed in Section D. At the start of the MARS forward-step algorithm for our example problem (step 24a), the initial subregion is again the entire domain i.e., $R_1 = D$. The single subregion MARS estimate of $f(\boldsymbol{x})$ is restricted to be identical to the recursive partitioning estimate,

$$\hat{f}(\boldsymbol{x}) \;=\; c_1 K_1(\boldsymbol{x}) \;=\; c_1 T_{0,r_1}(\boldsymbol{x}) \;=\; c_1 \;=\; \frac{1}{N}\sum_{1}^{N} y_i. \tag{25}$$

Again, let the exhaustive search in the first iteration of MARS identify the best partition of R_1 as $t^* = x_{2,25}$. Continuing, the three subregion MARS estimate of $f(\boldsymbol{x})$ obtained at the second step (first partition at $t^* = x_{2,25}$) is, with $T_{0,r_1}(\boldsymbol{x}) = 1$,

$$\begin{aligned} \hat{f}(\boldsymbol{x}) \;&= c_1\, K_1(\boldsymbol{x}) + c_2\, K_2(\boldsymbol{x}) + c_3\, K_3(\boldsymbol{x}) \\ &= c_1\, T_{0,r_1}(\boldsymbol{x}) + c_2\, T_{0,r_1}(\boldsymbol{x})\, T_{1,r_2}(\boldsymbol{x}) + c_3\, T_{0,r_1}(\boldsymbol{x})\, T_{1,r_3}(\boldsymbol{x}) \\ &= c_1 + c_2\,(t^* - x_2)_+ + c_3\,(x_2 - t^*)_+, \end{aligned} \tag{26}$$

$$\text{where } \boldsymbol{x} \in \begin{cases} R_1 & \text{if} \quad \boldsymbol{x} \in D \\ R_2 & \text{if} \quad x_2 < x_{2,25} \text{ and } \boldsymbol{x} \in R_1 \\ R_3 & \text{if} \quad x_2 \geq x_{2,25} \text{ and } \boldsymbol{x} \in R_1. \end{cases}$$

In the next iteration of the forward-step MARS algorithm the best partition point will occur within the subregions R_1, R_2 or R_3 and as in recursive partitioning, with one exception, will be chosen after evaluation of all potential partition points for each predictor variable within the three subregions. The exception, as discussed previously, prevents another partition on x_2 in R_2 or R_3 because it would create a truncated spline function of order greater than 1. With $M = 5$ the forward step of the MARS algorithm will be complete after a second partition in D. The final forward step MARS estimate of $f(\boldsymbol{x})$ for our example will include all terms in (26) and the additional two terms generated by the second partition. The model will have 5 single term product spline functions (excluding $T_{0,r_1}(\boldsymbol{x})$)

if the second partition occurs in R_1 while the model will have 3 single term product spline functions and two 2-way product spline functions if the second partition occurs in R_2 or R_3.

After the backward trimming procedure, the final MARS model retains the form of (22) with c_1 the coefficient of the product basis function $K_1(\boldsymbol{x})$ and the remaining terms the coefficients and product basis functions that survive the MARS backward step subregion deletion strategy. To provide an insight of predictor variable relationships we can rearrange the final MARS estimate of $f(\boldsymbol{x})$ in an ANOVA style decomposition (Friedman, 1988),

$$\hat{f}(\boldsymbol{x}) = c_1 + \sum_{V=1} c_j K_j(\boldsymbol{x}) + \sum_{V=2} c_j K_j(\boldsymbol{x}) + \ldots \tag{27}$$

where V indexes the number of truncated splines (excluding $T_{0,r_1}(\boldsymbol{x})$) in the product basis function $\{K_j(\boldsymbol{x})\}_{j=1}^{S}$. This method identifies any and all contributions to $\hat{f}(\boldsymbol{x})$ by variables of interest. Product basis functions with the index $V = 1$ reflect truncated linear trends and those with the index $V = 2$ reflect truncated 2-way interactions, etc. The ANOVA style decomposition (27) identifies which variables enter the model, whether they are purely additive, or are involved in interactions with other variables. Analysis of the ANOVA style decomposition facilitates interpretation of the MARS model.

F. NONLINEAR MODELING OF UNIVARIATE TIME SERIES USING MARS

As previously discussed in the introduction, most research in and applications of time series modeling and analysis is focused on linear models. This is due to the maturity of the theory for linear time series, and the numerous studies and statistical packages that exist to facilitate the use of linear time series models. However, more frequently than not, nonlinear time dependent systems abound that are not adequately handled by linear models. The use of linear models during the analysis of these nonlinear systems may require invalid assumptions that could lead to erroneous or misleading conclusions. For these systems we need to consider general classes of nonlinear models that readily adapt to the precise form of a nonlinear system of interest (Priestley, 1988 and Tong, 1985).

An example of a nonlinear time series system is that of sea-surface temperatures and the associated wind velocity and direction. Consider the sea-surface temperatures alone, a

specific example is the sea-surface temperatures analyzed by Breaker and Lewis (1985). A very clear nonlinearity in this time series is the abrupt, yearly spring transition to lower temperatures. The spring transition can be clearly seen in Figure 35 in Chapter IV, especially at about 2190 days. More particularly, the spring transition is strongly coupled with the wind direction, which shifts in the spring (Breaker and Lewis, 1988, pg 395). In addition there is an effect of the El Nino (a tropical warming) that occurs during some years. We return to this example in Chapter IV.

By letting the predictor variables for the τth value in a time series $\{X_\tau\}$ be $X_{\tau-1}$, $X_{\tau-2}$, ..., $X_{\tau-p}$, and combining these predictor variables into a linear additive function, one gets the well known linear AR(p) time series models (Priestley, 1988). What happens if we use the MARS methodology to model the effect of $X_{\tau-1}, X_{\tau-2}, \ldots, X_{\tau-p}$ on X_τ? The answer is that we still obtain autoregressive models. *However, these models can be nonlinear models in the sense that the lagged predictor variables can have threshold terms, in the form of truncated spline functions* (20) *and can also interact with the nonlinear terms formed with other lagged predictor variables.* The remainder of this chapter is a discussion of the form and analysis of these nonlinear univariate time series models.

Threshold time series models (models with partition points) are a class of nonlinear models that emerge naturally as a result of changing physical behavior. Within the domain of the predictor variables, different model forms are necessary to capture changes to the relationship between the predictor and response variables (a simple example of a threshold model is at equation (33)). Tong (1983) provides one threshold modeling methodology for this behavior (TAR – Threshold Autoregression) that identifies piecewise linear pieces of nonlinear functions over disjoint subregions of the domain D of the time series $\{X_\tau\}$, i.e., identify linear models within each disjoint subregion of the domain. One application of Tong's threshold modeling methodology is for nonlinear systems thought to possess periodic behavior in the form of stationary sustained oscillations (limit cycles). Tong's threshold methodology has tremendous power and flexibility for modeling of many times series. However, unless Tong's methodology is constrained to be continuous, it creates disjoint subregion models that are discontinuous at subregion boundaries.

With MARS, by letting the predictor variables be lagged values of a time series, one admits a *more general class* of continuous nonlinear threshold models than permitted by

Tong's TAR approach. The methodology for developing this class of nonlinear threshold models is called ASTAR (Adaptive Spline Threshold Autoregression). The fact that one obtains a more general class of continuous nonlinear threshold models can be shown using a simple example. Let X_τ for $\tau = 1, \ldots, N$, be a time series we wish to model with ASTAR using, for example, $p = 3$ lagged predictor variables namely, $X_{\tau-1}, X_{\tau-2}$ and $X_{\tau-3}$. Each forward step of the ASTAR algorithm selects *one and only one* set of new terms for the ASTAR model from the candidates specified by previously selected terms of the model. For our example problem the sets of candidates in the *initial* forward step of the ASTAR algorithm are

$$\begin{aligned}
&(X_{\tau-1} - t^*)_+ \text{ and } (t^* - X_{\tau-1})_+, \text{ or} \\
&(X_{\tau-2} - t^*)_+ \text{ and } (t^* - X_{\tau-2})_+, \text{ or} \\
&(X_{\tau-3} - t^*)_+ \text{ and } (t^* - X_{\tau-3})_+,
\end{aligned} \tag{28}$$

for some partition point (threshold) t^* in the individual domain of the lagged predictor variables. For our example problem, assume that ASTAR selects the lagged predictor variable $X_{\tau-2}$ with threshold value $t^* = t_1$ i.e., $(X_{\tau-2} - t_1)_+$ and $(t_1 - X_{\tau-2})_+$ are the initial terms (other than the constant) in the ASTAR model. The sets of candidates in the second forward step of the ASTAR algorithm includes *all candidates* in (28) and the new sets of candidates:

$$\begin{aligned}
&(X_{\tau-1} - t^*)_+(X_{\tau-2} - t_1)_+ \text{ and } (t^* - X_{\tau-1})_+(X_{\tau-2} - t_1)_+, \text{ or} \\
&(X_{\tau-3} - t^*)_+(X_{\tau-2} - t_1)_+ \text{ and } (t^* - X_{\tau-3})_+(X_{\tau-2} - t_1)_+, \text{ or} \\
&(X_{\tau-1} - t^*)_+(t_1 - X_{\tau-2})_+ \text{ and } (t^* - X_{\tau-1})_+(t_1 - X_{\tau-2})_+, \text{ or} \\
&(X_{\tau-3} - t^*)_+(t_1 - X_{\tau-2})_+ \text{ and } (t^* - X_{\tau-3})_+(t_1 - X_{\tau-2})_+,
\end{aligned} \tag{29}$$

due to the initial selection of $(X_{\tau-2} - t_1)_+$ and $(t_1 - X_{\tau-2})_+$, and where t^* is to be determined. Thus one could have multiple thresholds on one variable, say $X_{\tau-2}$, by again selecting as the next set of model terms $(X_{\tau-2} - t^*)_+$ and $(t^* - X_{\tau-2})_+$ from (28) for some partition point $t^* \neq t_1$. The sets of candidates for each subsequent forward step of the

ASTAR algorithm is nondecreasing in size and is based on previously selected terms of the model. As discussed in Section D, the forward-step algorithm is followed by a backward-step algorithm that trims the excess $(M - S)$ terms from the model, where S is the final number of terms in the model, with $1 \leq S \leq M$.

Let the predictor variables in MARS for the τth value in a time series $\{X_\tau\}$ be $X_{\tau-1}$, $X_{\tau-2}$, ..., $X_{\tau-p}$, which we represent as $\boldsymbol{X}_{\tau-1}^p$. Following (22), the functional form of the ASTAR model that estimates X_τ is

$$\hat{X}_\tau = \sum_{j=1}^{S} c_j K_j(\boldsymbol{X}_{\tau-1}^p), \tag{30}$$

where $\hat{X}_\tau$ is an additive function of the product spline basis functions $\{K_j(\boldsymbol{X}_{\tau-1}^p)\}_{j=1}^S$ associated with the subregions $\{R_j\}_{j=1}^S$. The functional form of the ASTAR model (30) may be expanded using the ordered sequences of truncated spline functions (20 and 21) that define each product spline basis function. Let a and b be dummy variables that index the ordered sequence of truncated spline functions $T_{a,\boldsymbol{r}_b}(\boldsymbol{X}_{\tau-1}^p)$ such that $0 \leq a < b \leq j$. The functional form of the ASTAR model (30) for the τth value in a time series $\{X_\tau\}$ using this expansion is

$$\hat{X}_\tau = \sum_{j=1}^{S} c_j \prod_{T_{a,\boldsymbol{r}_b} \in K_j} [sgn_v(X_{\tau-v} - t)]_+ \tag{31}$$

where the argument, $\boldsymbol{X}_{\tau-1}^p$, of $T_{a,\boldsymbol{r}_b}(\boldsymbol{X}_{\tau-1}^p)$, and $K_j(\boldsymbol{X}_{\tau-1}^p)$ is suppressed for simplicity. Also $\boldsymbol{r}_b = (\pm v, t)$ from (20), and sgn_v is the sign of v that determines a left $(-v)$ or right $(+v)$ truncated spline function.

By modeling univariate time series using ASTAR we overcome some of the limitations of Tong's approach. The ASTAR methodology creates threshold time series models that are naturally continuous in the domain of the predictor variables, and it allows interactions among lagged predictor variables. Also, the ASTAR time series model can have multiple lagged predictor variable thresholds, e.g., the model (29) if the new partition point $t^* \neq t_1$. In contrast, Tong's methodology creates threshold models from piecewise linear models whose terms are restricted to the initial sets of candidates of the ASTAR algorithm (equation (28) for our example). Tong's threshold models do not allow interactions among lagged

predictor variables and are usually limited to a single threshold value over all the lagged predictor variables because of the difficulties associated with the threshold selection process.

An initial question that exists is whether MARS is able to identify and model simple linear and nonlinear times series models? If not it would be of little value to use MARS with real data with unknown structure. In the next two sections, simulation experiments are used to determine the ability of MARS to detect and model simple linear and nonlinear time series models. The simulation of an AR(1) model with known coefficients examines the ability of ASTAR to detect and model a simple *linear* time series. The simulation of a threshold model with 'AR(1)-like' models in each disjoint subregion examines the ability of ASTAR to detect and model simple *nonlinear* threshold time series. The interest in these simulations is two-fold: how often was the true model identified, and if so, how well were the parameters K and ρ estimated. Finally, as a demonstration of the ability of ASTAR to model a real univariate time series system, the last section of this chapter considers the widely studied yearly Wolf sunspot numbers, a nonlinear time series with periodic behavior.

1. AR(1) Time Series Model Simulations

The initial simulation experiments are of the first order autoregressive (AR(1)) time series model,

$$X_\tau = \rho X_{\tau-1} + K + \epsilon_\tau \tag{32}$$

where $\tau = 1, 2, \ldots, N$ indexes the time series, ρ is a constant coefficient varied within experiments, K is the model constant, taken to be zero, and ϵ_τ is $N(0, \sigma_\epsilon^2)$. The model is usually considered under the stationarity conditions ($| \rho | < 1$), but non-stationary processes such as random walks ($| \rho | = 1$) and explosive processes ($| \rho | > 1$), are also of interest.

Two categories of experiments were conducted using the AR(1) time series model.

The first experiment required ASTAR to identify and estimate parameters of a time series model from the simulated data of the AR(1) time series model using one lag predictor variable $X_{\tau-1}$, and using $M = 3$, the maximum number of subregions in the forward-step ASTAR procedure. The first experiment's alternative models (to the AR(1) time series model) either have no $X_{\tau-1}$ term (a constant model) or have a $X_{\tau-1}$ term with

a threshold value t greater than $\min\{X_{\tau-1}\}_{\tau=1}^{N-1}$. In this case we call the threshold value t an internal threshold.

The second experiment required ASTAR to identify and estimate parameters of a time series model from the simulated data of the AR(1) time series model when up to four lag predictor variables, $\{X_{\tau-i}\}_{i=1}^{4}$, are allowed, and using $M = 8$, the maximum number of subregions allowed in the forward-step ASTAR procedure. The second experiment's alternative models include constant models, time series models with an internal threshold value, or any time series model that includes a term other than $X_{\tau-1}$.

Several simulation results are shown in Figures 2 – 7 for $\rho = .5, .7$ and $.9$, $K = 0$, and $\epsilon_\tau = \mathrm{N}(0,1)$. Each figure is a series of box plots for the estimated coefficients of the 100 simulated models correctly identified as AR(1) time series models by ASTAR for increasing values of N, the statistics for $\hat{\rho}$ being given in the top set of boxplots, and the statistics for $\widehat{K}$ in the bottom set. The true value of each model coefficient is identified by the dashed line across the box plots. At the top of each figure we see the length N of each simulated time series, the number C of the 100 simulated models correctly identified by the ASTAR procedure, and the equivalent sample size for independent data, $Eq\ S\ SIZE = (N/\sum_{i=-\infty}^{\infty}\rho^i)$ (Priestley, 1981). Underneath each box plot is summary information for the coefficient estimates of the correctly identified AR(1) time series models i.e., the sample mean and sample standard deviation of the estimated values in the box plots. By comparing the true and the estimated values of the model coefficients across increasing values of N it is observed that the estimated values of the coefficients tend to the true value as N increases. Also, in all but one simulation the number of correctly identified models C rises to 100 for increasing values of N. Note that the ASTAR estimates for ρ have negative bias for small values of N that generally decreases as N increases. The downward bias of $\hat{\rho}$ is similar to that identified by Kendall et al. (1983) and others when using data for estimating autocorrelations.

2. Nonlinear Threshold Time Series Model Simulations

To observe the ability of ASTAR to capture nonlinear threshold time series model characteristics, simulation experiments of the 2-subregion threshold time series model (Tong,

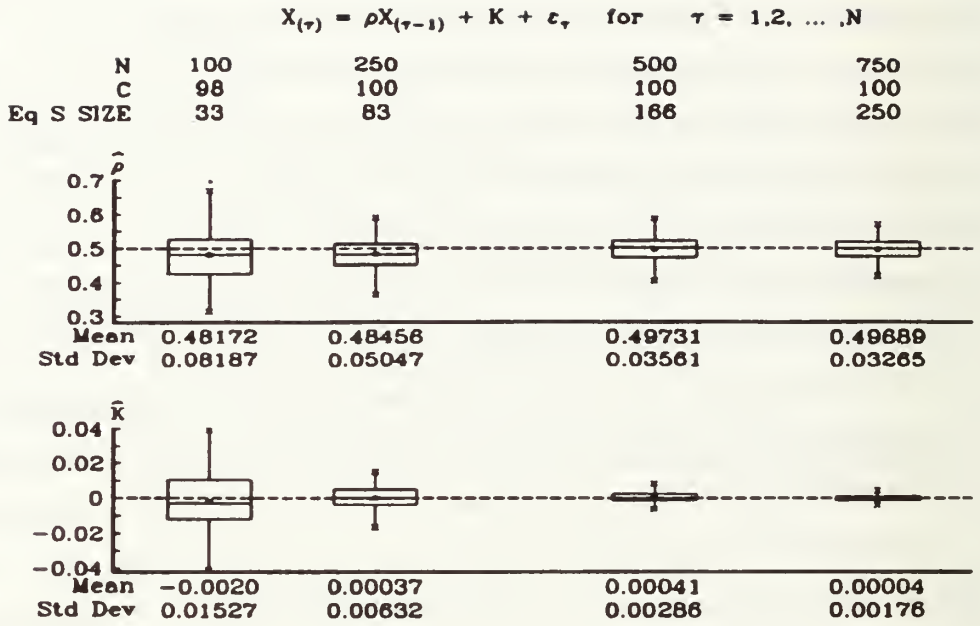

Figure 2. AR(1) MODEL SIMULATION: ASTAR estimates for $\rho = .5, K = 0$ using $\sigma_\epsilon^2 = N(0,1)$ from C simulations of an AR(1) model for increasing values of N, with $P = 1$ lag predictor variables, and $M = 3$, the number of forward-step subregions permitted in the ASTAR algorithm. Each simulation consists of 100 replications. The boxplots are for the estimates $\hat{\rho}$ and $\widehat{K}$ of the model parameters when ASTAR correctly identified the AR(1) model. *For $N = 100$, 2 simulations were incorrectly identified as constant models.*

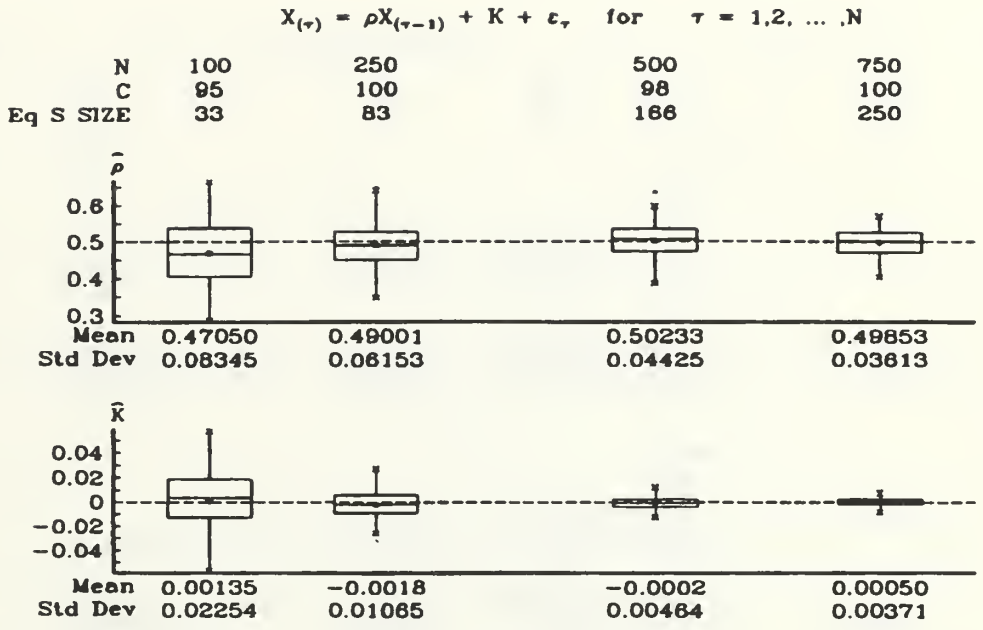

Figure 3. AR(1) MODEL SIMULATION: ASTAR estimates for $\rho = .5, K = 0$ using $\sigma_\epsilon^2 = N(0,1)$ from C simulations of an AR(1) model for increasing values of N with $P = 4$ lag predictor variables, and $M = 8$, the number of forward-step subregions permitted in the ASTAR algorithm. Each simulation consists of 100 replications. The boxplots are for the estimates $\hat{\rho}$ and $\widehat{K}$ of the model parameters when ASTAR correctly identified the AR(1) model. *For $N = 100$, 5 simulations were incorrectly identified as; 2 constant models, 1 AR(2) model and 2 AR(3) models. For $N = 500$, 2 simulations were incorrectly identified as constant models.*

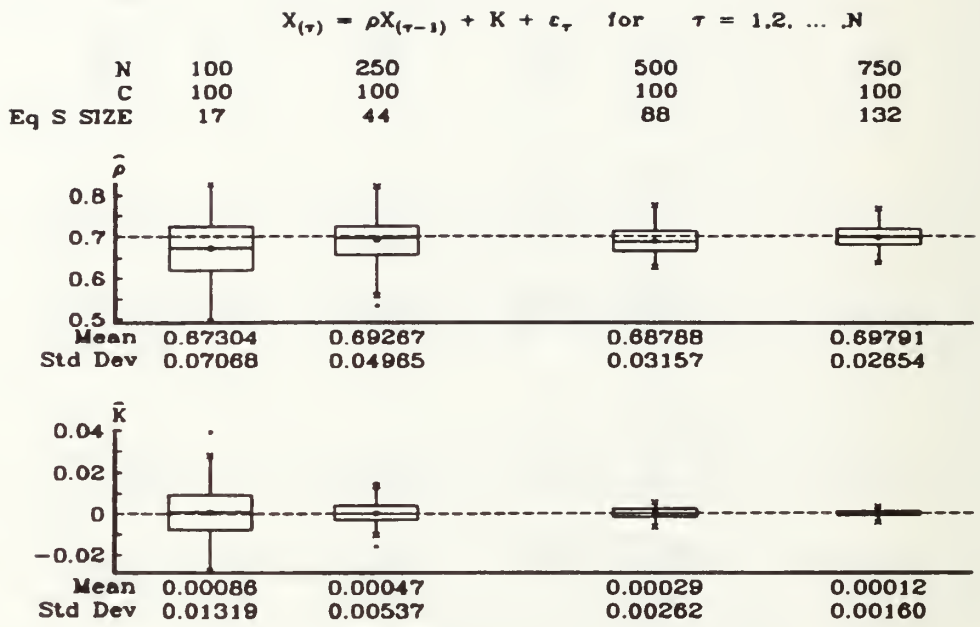

Figure 4. AR(1) MODEL SIMULATION: ASTAR estimates for $\rho = .7, K = 0$ using $\sigma_\epsilon^2 = N(0,1)$ from C simulations of an AR(1) model for increasing values of N, with $P = 1$ lag predictor variables, and $M = 3$, the number of forward-step subregions permitted in the ASTAR algorithm. Each simulation consists of 100 replications. The boxplots are for the estimates $\hat{\rho}$ and $\widehat{K}$ of the model parameters when ASTAR correctly identified the AR(1) model. *Here all cases were correctly identified.*

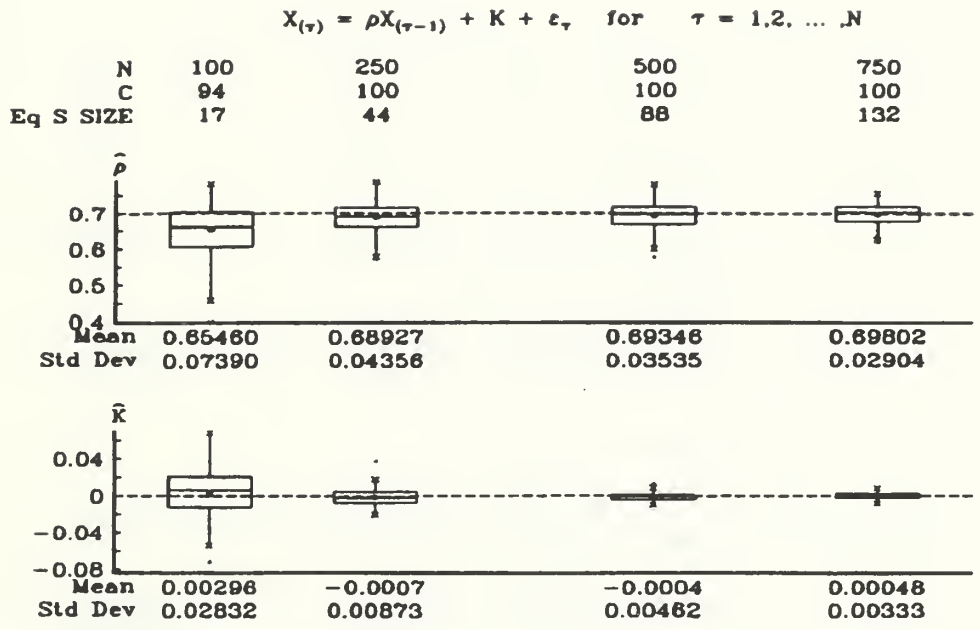

Figure 5. AR(1) MODEL SIMULATION: ASTAR estimates for $\rho = .7, K = 0$ using $\sigma_\epsilon^2 = N(0,1)$ from C simulations of an AR(1) model for increasing values of N with $P = 4$ lag predictor variables, and $M = 8$, the number of forward-step subregions permitted in the ASTAR algorithm. Each simulation consists of 100 replications. The boxplots are for the estimates $\hat{\rho}$ and $\widehat{K}$ of the model parameters when ASTAR correctly identified the AR(1) model. *For $N = 100$, 6 simulations were incorrectly identified as; 2 AR(2) models, 2 AR(3) model and 2 AR(4) models.*

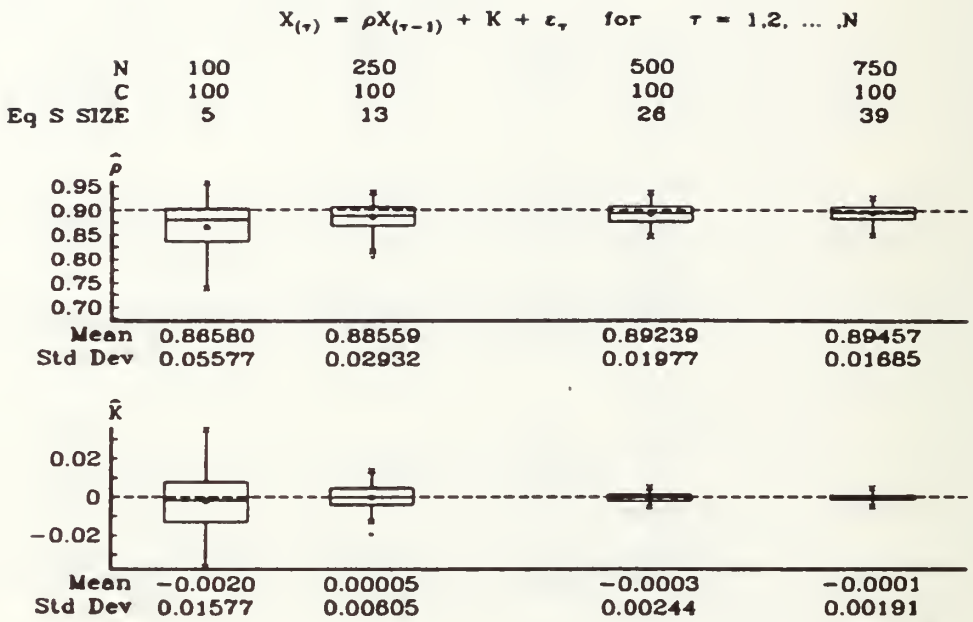

Figure 6. AR(1) MODEL SIMULATION: ASTAR estimates for $\rho = .9, K = 0$ using $\sigma_\epsilon^2 = N(0,1)$ from C simulations of an AR(1) model for increasing values of N, with $P = 1$ lag predictor variables, and $M = 3$, the number of forward-step subregions permitted in the ASTAR algorithm. Each simulation consists of 100 replications. The boxplots are for the estimates $\hat{\rho}$ and $\widehat{K}$ of the model parameters when ASTAR correctly identified the AR(1) model. *Here all cases were correctly identified.*

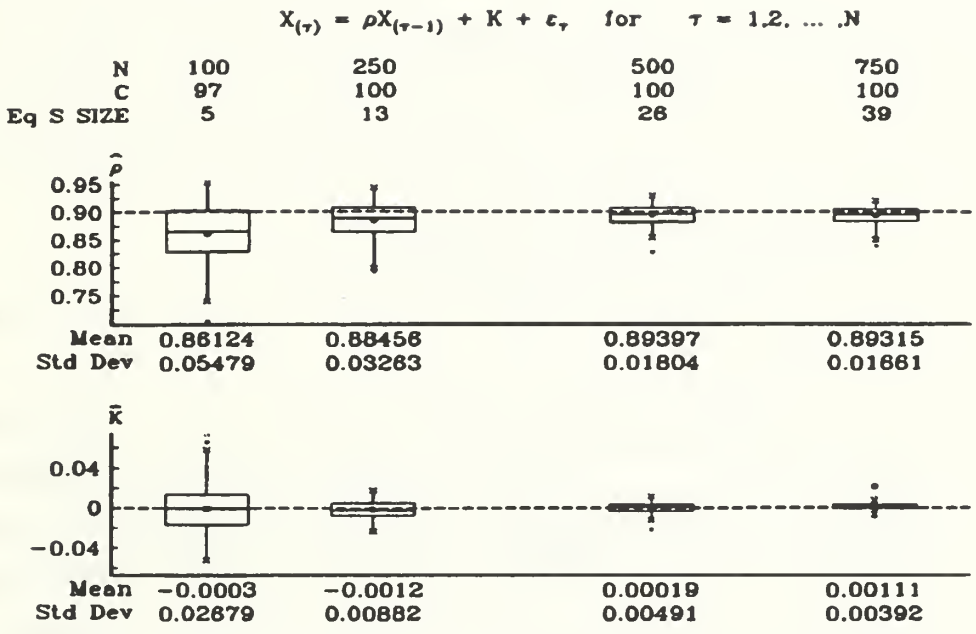

Figure 7. AR(1) MODEL SIMULATION: ASTAR estimates for $\rho = .9, K = 0$ using $\sigma_\epsilon^2 = N(0,1)$ from C simulations of an AR(1) model for increasing values of N with $P = 4$ lag predictor variables, and $M = 8$, the number of forward-step subregions permitted in the ASTAR algorithm. Each simulation consists of 100 replications. The boxplots are for the estimates $\hat{\rho}$ and $\widehat{K}$ of the model parameters when ASTAR correctly identified the AR(1) model. *For $N = 100$, 3 simulations were incorrectly identified as; 2 AR(2) models and 1 AR(3) model.*

1983),

$$X_\tau = \begin{cases} \rho_1 X_{\tau-1} + \epsilon_\tau & \text{if } X_{\tau-1} \leq 0 \\ \rho_2 X_{\tau-1} + \epsilon_\tau & \text{if } X_{\tau-1} > 0 \end{cases} \tag{33}$$

were considered, where $\tau = 1, 2, \ldots, N$ indexes the time series, ρ_1 and ρ_2 are constant coefficients varied for different experiments and ϵ is $N(0, \sigma_\epsilon^2)$. This is the simplest threshold model which has been proposed and provides a convenient starting point for initial evaluation and validation of the ASTAR procedure. Note that the nonlinear threshold time series model (33) has an 'AR(1)-like' model in each subregion, which implies that X_τ can have different variance in each of the two subregions since the variance of ϵ is assumed constant in each region. Also for a threshold at $X_{\tau-1} = 0$, the expected number of sample values in each subregion will be the same only if $\rho_1 = -\rho_2$.

Two categories of experiments were conducted using the nonlinear threshold time series model.

The first experiment required ASTAR to identify and estimate parameters of a time series model from the simulated data of the nonlinear threshold time series model using one lag predictor variable $X_{\tau-1}$, and using $M = 3$, the maximum number of subregions in the forward-step ASTAR procedure. The first experiment's alternative models include the constant model, linear AR(1) time series models, or nonlinear time series models that have more than one internal threshold.

The second experiment required ASTAR to identify and estimate parameters of a time series model from the simulated data of the nonlinear threshold time series model when up to four lag predictor variables, $\{X_{\tau-i}\}_{i=1}^{4}$, are allowed, and using $M = 10$, the maximum number of subregions allowed in the forward-step ASTAR procedure. The second experiment's alternative models include the constant model, linear and nonlinear time series models with terms other than $X_{\tau-1}$, or nonlinear time series models with more than one internal threshold value on $X_{\tau-1}$.

Several simulation results are shown in Figures 8 – 11 for $\rho_1, \rho_2 = .7, .3$ and $-.6, .6$, and $\epsilon_\tau = \text{N}(0,.25)$. As with the previous AR(1) time series model simulation experiments, each figure is a series of box plots for the estimated coefficients of the 100 simulated models correctly identified as nonlinear threshold time series models by ASTAR for increasing values

of N. The true value of each model coefficient is identified by the dashed line across the box plots. At the top of each figure is the length N of each simulated time series, and the number C of the 100 simulated models correctly identified by the ASTAR procedure. Underneath each box plot is summary information for the coefficient estimates of the correctly identified nonlinear threshold time series models i.e., the sample mean and sample standard deviation of the estimated values in the boxplots. Note that the number of correctly identified models rises for increasing values of N. However, a consistent improvement in the mean and standard deviation for the estimated values of the model coefficients is not always observed for increasing values of N. For the most part this is attributed to the increasing number of correctly identified models for increasing values of N.

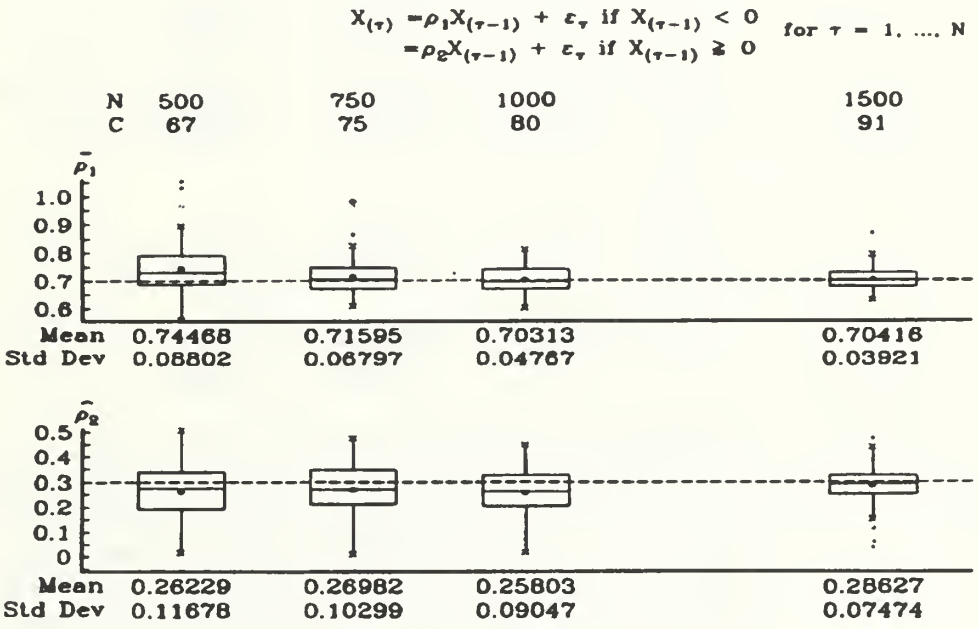

Figure 8. THRESHOLD MODEL SIMULATION: ASTAR model estimates for $\rho_1, \rho_2 = .7, .3$ using $\sigma_\epsilon^2 = N(0, .25)$ from C simulations of a threshold model for increasing values of N, with $P = 1$ lag predictor variables, and $M = 3$, the number of forward-step subregions permitted in the ASTAR algorithm. Each simulation consists of 100 replications. The boxplots are for the estimates $\hat{\rho}_1$ and $\hat{\rho}_2$ of the model parameters when ASTAR correctly identified the threshold model. *The models of the simulations that ASTAR did not correctly identify as the threshold model contained an incorrect number of subregions or lacked an AR(1) term in one of the two subregions.*

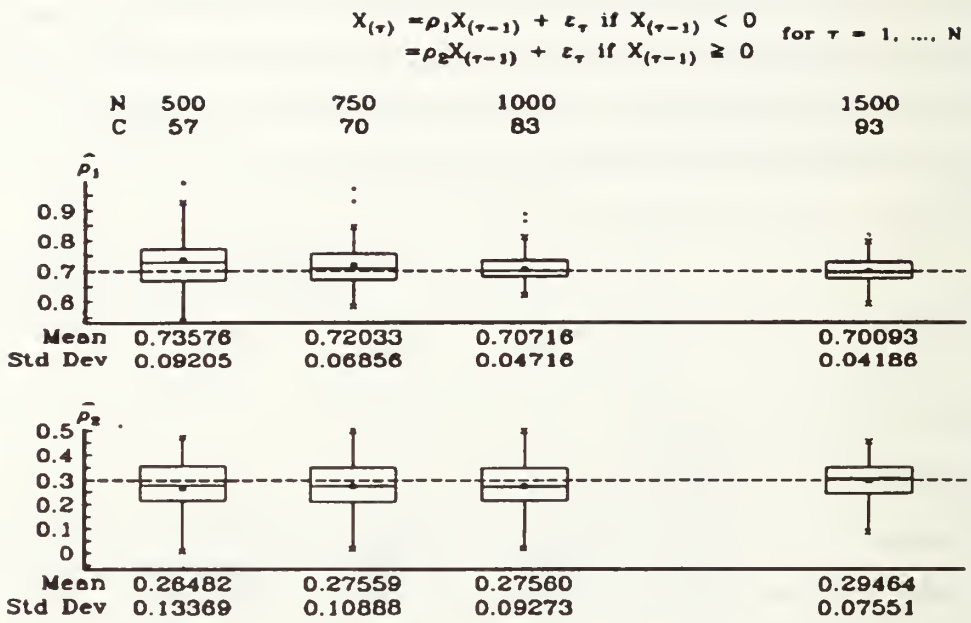

Figure 9. THRESHOLD MODEL SIMULATION: ASTAR estimates for $\rho_1, \rho_2 = .7, .3$ using $\sigma_\epsilon^2 = N(0, .25)$ from C simulations of a threshold model for increasing values of N, with $P = 4$ lag predictor variables, and $M = 10$, the number of forward-step subregions permitted in the ASTAR algorithm. Each simulation consists of 100 replications. The boxplots are for the estimates $\hat{\rho}_1$ and $\hat{\rho}_2$ of the model parameters when ASTAR correctly identified the threshold model. *The models of the simulations that ASTAR did not correctly identify as the threshold model contained an incorrect number of subregions, lacked an AR(1) term in one of the two subregions or contained terms with $X_{\tau-2}, X_{\tau-3}$, or $X_{\tau-4}$.*

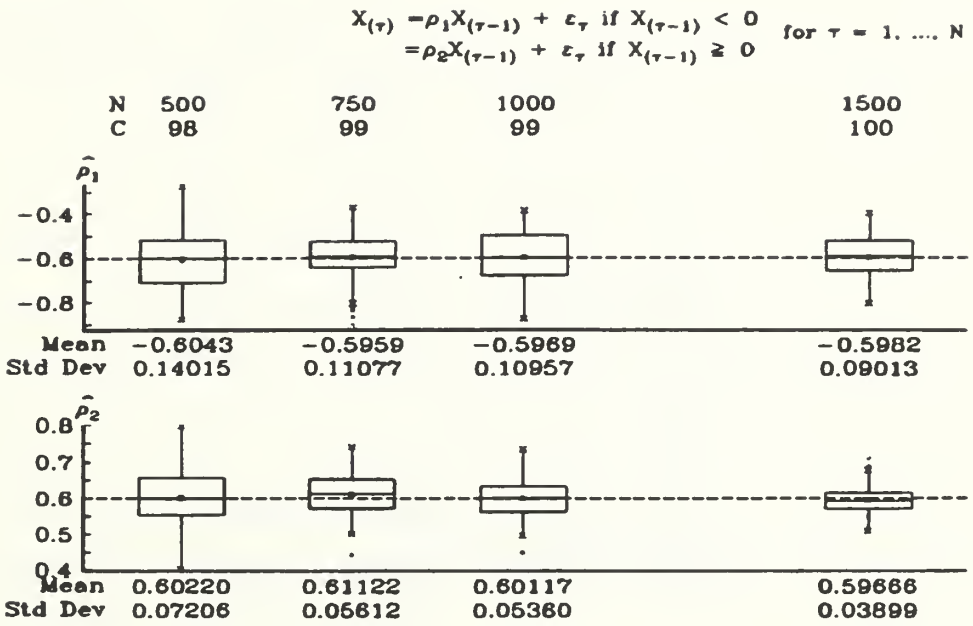

Figure 10. THRESHOLD MODEL SIMULATION: ASTAR estimates for $\rho_1, \rho_2 = -.6, .6$ using $\sigma_\epsilon^2 = N(0, .25)$ from C simulations of a threshold model for increasing values of N, with $P = 1$ lag predictor variables, and $M = 3$, the number of forward-step subregions permitted in the ASTAR algorithm. Each simulation consists of 100 replications. The boxplots are for the estimates $\hat{\rho}_1$ and $\hat{\rho}_2$ of the model parameters when ASTAR correctly identified the threshold model *The models of the simulations that ASTAR did not correctly identify as the threshold model contained an incorrect number of subregions or lacked an AR(1) term in one of the two subregions.*

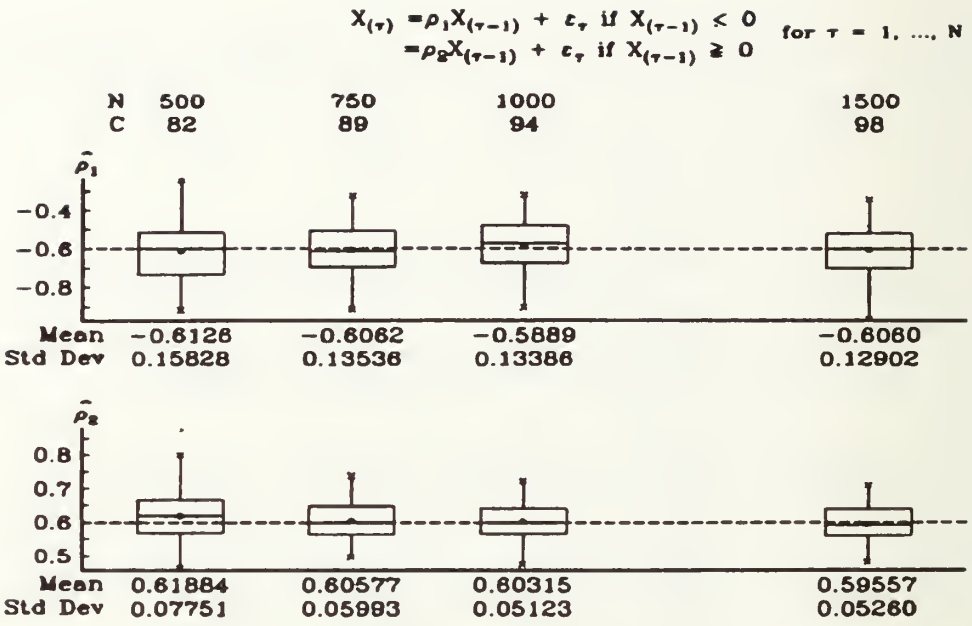

Figure 11. THRESHOLD MODEL SIMULATION: ASTAR estimates for $\rho_1, \rho_2 = -.6, .6$ using $\sigma_\epsilon^2 = N(0, .25)$ from C simulations of a threshold model for increasing values of N, with $P = 4$ lag predictor variables, and $M = 10$, the number of forward-step subregions permitted in the ASTAR algorithm. Each simulation consists of 100 replications. The boxplots are for the estimates $\hat{\rho}_1$ and $\hat{\rho}_2$ of the model parameters when ASTAR correctly identified the threshold models *The models of the simulations that ASTAR did not correctly identify as the threshold model contained an incorrect number of subregions, lacked an AR(1) term in one of the two subregions or contained terms with* $X_{\tau-2}, X_{\tau-3}$, *or* $X_{\tau-4}$.

3. Threshold Modeling of the Yearly Wolf Sunspot Numbers

As an illustration of ASTAR's ability to model an actual time series we examined 221 (1700-1920) of the yearly Wolf sunspot numbers. These yearly Wolf sunspot numbers are relative measures of the average monthly sunspot activity on the surface of the sun (see, e.g., Scientific American, February 1990). The analysis was performed on the yearly sunspot numbers to facilitate comparison of the MARS methodology with other nonlinear time series modeling efforts (analysis of monthly sunspot numbers would also be of interest). Some of the early analysis and modeling of the yearly sunspot numbers was performed by Yule (1927) as an example for introducing autoregressive models. Recently suggested nonlinear models of the yearly sunspot numbers include threshold models (Tong, 1983, 1985) and bilinear models (Rao and Gabr, 1984). A detailed review of the history of the sunspot numbers is provided by Izenman (1983).

The data (Figure 12) is quite 'periodic' but has nonsymmetric cycles with extremely sharp peaks and troughs. The cycles (Table 1) generally vary between 10 and 12 years with the greater number of sunspots concentrated in each descent period versus the accompanying ascent period. The average ascent period is 4.60 years and the average descent period is 6.58 years. Attempts to model the data with a fixed cycle period signal plus (possibly correlated) noise have failed because the cyclical component in the spectrum (Figure 14, top) is quite spread out and diffuse.

TABLE 1. ASCENT AND DESCENT PERIODS OF THE YEARLY WOLF SUNSPOT NUMBERS (1700-1920).

Ascent period	5	5	4	5	6	6	3	3	3	6	6	
Descent period	7	6	6	6	5	5	6	6	11	6	7	
Ascent (cont)	7	4	5	4	3	5	4	4	4			
Descent (cont)	3	6	8	7	8	6	8	8				

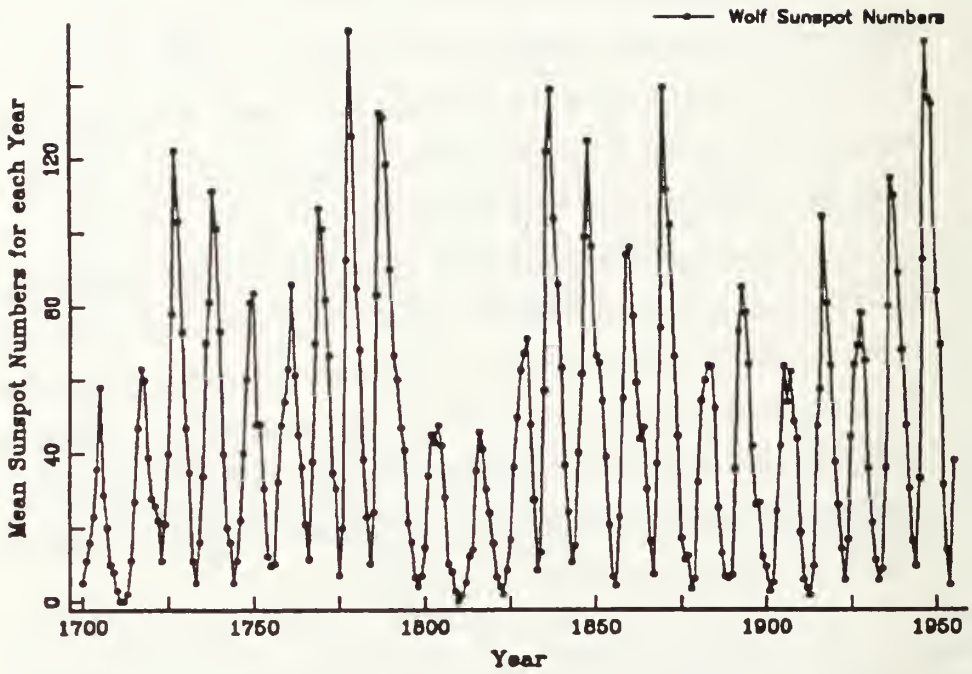

Figure 12. The yearly Wolf sunspot numbers (1700-1955). The data is quite 'periodic' but has nonsymmetric cycles with extremely sharp peaks and troughs. The cycles generally vary between 10 and 12 years with the greater number of sunspots concentrated in each descent period versus the accompanying ascent period.

a. *Stable Periodic Limit Points in Threshold Models*

One of the interesting characteristics of Tong's analysis of the yearly sunspot numbers included the development of threshold models with stationary harmonic behavior or limit cycles, i.e., models in which $f(\boldsymbol{X})$ is a deterministic threshold function with a limit cycle, perturbed by Gaussian white noise. Using Tong (1983), let $\tau = 1, 2, \ldots$ index a times series and let $\boldsymbol{X}_\tau^p = \{X_\tau, X_{\tau-1}, \ldots, X_{\tau-p+1}\}$ denote a p-dimensional vector in $D \in \Re^p$ that satisfies the equation,

$$\boldsymbol{X}_\tau^p = f(\boldsymbol{X}_{\tau-1}^p), \tag{34}$$

where f is a vector-valued function. Let $f^j(\boldsymbol{X})$ denote the jth iterate of f, i.e.,

$$f^j(\boldsymbol{X}) = \underbrace{f(f(f(\ldots(f(\boldsymbol{x}))\ldots)))}_{\text{j of them}}. \tag{35}$$

We say that a p-dimensional vector $\boldsymbol{X}^{*p}$ is a *stable limit point* of the function f with respect to the domain D if

$$f^j(\boldsymbol{X}_0) \rightarrow \boldsymbol{X}^{*p} \text{ as } j \rightarrow \infty \quad \forall\ \boldsymbol{X}_0 \in D. \tag{36}$$

Also, we say that a p-dimensional vector $\boldsymbol{C}_1^p$ is a *stable periodic limit point* with period $T > 1$ of the function f with respect to the domain D if

$$f^{jT}(\boldsymbol{X}_0) \rightarrow \boldsymbol{C}_1^p \text{ as } j \rightarrow \infty \quad \forall\ \boldsymbol{X}_0 \in D, \tag{37}$$

and the convergence does not hold for any divisor of T. It follows that $\boldsymbol{C}_1^p$, $f^1(\boldsymbol{C}_1^p)$, $f^2(\boldsymbol{C}_1^p), \ldots, f^{T-1}(\boldsymbol{C}_1^p)$ are simultaneously distinct stable periodic points of the function f with respect to D. If we let $f^i(\boldsymbol{C}_1^p)$ be denoted by $\boldsymbol{C}_{i+1}^p, i = 0, 1, \ldots, T-1$, then the set $\{\boldsymbol{C}_1^p, \boldsymbol{C}_2^p, \boldsymbol{C}_3^p, \ldots, \boldsymbol{C}_{T-1}^p\}$ is called a stable limit cycle of the function f with respect to D.

b. *ASTAR Models for the Yearly Wolf Sunspot Numbers*

The primary interest in limit cycles is for investigating the underlying characteristics of the true time series function $f(\boldsymbol{X})$ given at (1). If the cyclical behavior of $f(\boldsymbol{X})$ for the yearly sunspot numbers can be modeled as a limit cycle perturbed by Gaussian white noise, then when applying ASTAR to the yearly sunspot numbers it would be

satisfying to identify an underlying limit cycle in the estimate of $f(\boldsymbol{X})$. With this objective in mind 20 ASTAR models of the yearly sunspot numbers were investigated. The period of the modeling effort (1700-1920) corresponds to similar modeling efforts by Tong (1983, 1985) and Rao and Gabr (1984). The maximum order of each model (number of lagged predictor variables) was restricted to 20 and the first 20 sunspot numbers (1700-1719) were used for model initialization. It might be noted that the ASTAR models were identified with MARS 2.0 installed on an IBM3033 Computer using VS Fortran. Each model required from 15 to 30 seconds of CPU time.

Table 2 provides a summary of the 20 ASTAR models for the yearly sunspot numbers (1720-1920), ordered by the mean sum of squares (MSS) of the fitted residuals for each ASTAR model. The first three columns identify, respectively, the model number, MSS and the modified generalized cross validation criterion GCV^* given in (23). The fourth through sixth columns identify the number of estimated parameters, the number of partition points and the maximum level of interaction in each model. Columns seven and eight identify the length (in years) of each model's limit cycle (if one exists) and the number and lengths (in years) of the one or more type 'subcycles' (ascent and descent periods) within the limit cycle. We use MSS instead of $\text{MSS}^{1/2}$ to facilitate comparison of the ASTAR models with other modeling efforts of the yearly sunspot numbers.

The different models in Table 2 occurred because the user parameters of the ASTAR algorithm were varied. These parameters include: MI = 2, 3, or 4, the maximum level of lagged predictor variable interaction; MS = 10 and 18, the minimum separation of a lagged predictor variable's partition points; M = 15, 20, and 25, the number of steps during the forward-step algorithm; and p = 12 or 18, the number of lagged predictor variables (12 lagged predictor variables correspond to the maximum order of the model used by Tong (1985) for prediction of the sunspot numbers). The separation of a predictor variable's partition points may be thought of as a smoothing parameter similar to the bandwidth in kernel smoothing. Some of the resulting models were identical. For example, identical 3-way interaction models could result from using the different model parameters MI = 3 and MI = 4 if the MARS algorithm rejects all 4-way interactions. Also, most of the models formed using MI = 2 were not of interest. Note that Chapter IV provides a discussion of the user parameters within the MARS algorithm.

TABLE 2. ASTAR MODELS FOR THE YEARLY WOLF SUNSPOT NUMBERS (1720-1920).

	MSS	*GCV**	Number of Model Parameters	Number of Interior Thresholds	Level of Model Interaction	Length of Limit Cycle (in years)	Number (Lengths) of Subcycles
1	91.4	151.7	25	9	3	—	
2	91.6	136.4	16	4	4	225	27 (8,9)
3	95.3	157.9	18	4	5	—	
4	101.0	130.4	15	6	4	43	4 (10,11)
5	103.6	183.9	18	3	4	—	
6	110.5	187.6	17	3	3	167	15 (11,12)
7	111.7	153.9	14	4	3	9	1 (9)
8	113.0	162.8	19	7	2	9	1 (9)
9	114.1	141.0	14	6	3	137	13 (10,11)
10	114.2	160.8	14	3	4	—	
11	114.2	194.7	17	3	3	120	11 (10,11)
12	115.9	162.9	13	3	3	—	
13	115.9	163.6	13	3	4	120	11 (10,11)
14	116.0	174.3	13	2	4	94	10 (9,10)
15	117.6	190.9	15	2	4	—	
16	119.5	171.2	14	3	3	—	
17	119.6	206.6	18	3	3	23	4 (5,6)
18	119.8	164.4	11	2	3	133	12 (11,12)
19	125.6	172.7	11	2	2	78	7 (11,12)
20	126.2	192.7	13	1	3	—	

Some form of a limit cycle exists in 12 of the 20 ASTAR models. Also, 7 of the 12 models, namely 4,6,9,11,13,18 and 19, provide limit cycles with lengths 43, 167, 137, 120, 120, 133 and 78 respectively, and 'subcycles' with lengths and range similar enough to the behavior of the yearly sunspot data (Table 1) to warrant further analysis. Of these 7 models, 2 (Models 4 and 9) have both low GCV^* values and provide fitted residuals that appear, using test statistics and graphical analysis, e.g., Q-Q plots, to be independent and Gaussian. Some of the test statistics for the fitted residuals of these two models are provided in Table 3.

TABLE 3. STATISTICS FOR THE FITTED RESIDUALS OF ASTAR MODELS 4 AND 9 OF THE YEARLY WOLF SUNSPOT NUMBERS (1720-1920).

	Model 4	Model 9	
Mean	0.000	0.000	
GCV^*	130.4	141.0	
Skewness	.346	0.0813	0 for normal distribution
Kurtosis	0.153	0.673	0 for normal distribution
K-S	.349	.275	level of significance
C-M	$> .15$	$> .15$	level of significance
A-D	$> .15$	$> .15$	level of significance
L-M	.0466	.6892	level of significance

The Skewness and Kurtosis statistics serve as a general indicator of the symmetry and heaviness of the tails for the sample distribution function of the fitted residuals $\widehat{F}_{\epsilon}(x)$. The Kolmogorov-Smirnov (K-S) test statistic measures the maximum absolute distance between $\widehat{F}_{\epsilon}(x)$ and the hypothesized true normal N(0,1) distribution function $F_{\mathbf{X}}(x)$ while the Cramer-von Mises (C-M) statistic measures the integral of the squared distance between the two functions. A drawback to the K-S and C-M tests are that they lack sensitivity to departures from the null hypothesis that occur in the tails of a distribution. As an approach to overcome the lack of sensitivity of the K-S and C-M tests, the Anderson-Darling

(A-D) test statistic weights the distances between the two functions. A final test for independent and Gaussian error structure is provided by the Lin-Mudhoekar (L-M) (1980) test statistic which tests for asymmetry. Even though the GCV^* for Model 4 is lower than that for Model 9, we rejected Model 4 due to the low level of significance of the L-M test statistic and identified Model 9 as the best model (with limit cycle) of the 20 models considered in the initial analysis.

Note that the MARS algorithm generated ASTAR Model 9 using 20 lagged predictor variables that were permitted to form 1, 2, and 3-way interactions during a maximum of $M = 15$ forward steps of the forward-step algorithm. The minimum span between threshold values for a single predictor variable was 18 sunspots. This span was chosen because there were 18 sunspot cycles between 1720 and 1920. Model 9 is

ASTAR Model 9

$$\hat{X}_\tau = \begin{cases} 2.711 + .960X_{\tau-1} + .332(47.0 - X_{\tau-5})_+ - .257(59.1 - X_{\tau-9})_+ \\ \\ - .003X_{\tau-1}(X_{\tau-2} - 26.0)_+ + .017X_{\tau-1}(44.0 - X_{\tau-3})_+ \\ \\ - .032X_{\tau-1}(17.1 - X_{\tau-4})_+ + .004X_{\tau-1}(26.0 - X_{\tau-2})_+(X_{\tau-5} - 41.0)_+ \end{cases} \tag{38}$$

where $(x)_+$ is a plus function with value x if $x > 0$ and 0 otherwise. Model 9 has 14 parameters with 8 terms (a constant term and 3 one-way, 3 two-way and 1 three-way interactions) and 6 threshold values (1 each on $X_{\tau-2}$, $X_{\tau-3}$, $X_{\tau-4}$, and $X_{\tau-9}$ and 2 on $X_{\tau-5}$).

Figures 13-19 are various plots of the fitted values and residuals of ASTAR Model 9. Figure 13 shows the fitted values of the model versus the yearly Wolf sunspot numbers (1720-1920). The model fit is further examined using the estimated normalized periodogram (Figure 14) of the sunspot number data [top] and model fit [bottom], empirical quantile-quantile plot (Figure 15) and autocorrelation function plots (Figure 16) of the fitted values of the model versus the yearly Wolf sunspot numbers (1720-1920). The model appears to equally overfit and underfit the peaks and troughs as it captures the general structure

of the yearly sunspot numbers. Again, note the spread of the cyclical component in the spectrum (Figure 14) that has complicated efforts to model the sunspot numbers with a fixed period signal plus (possibly correlated) noise. The fitted residuals of the model are examined using residual versus time and fit plots (Figure 17) and the residual autocorrelation function plot (Figure 18). In Figure 17 the slight lack of negative residuals for small fitted values of the model is attributed to the yearly sunspot numbers being positive random variables. In Figure 18 no pattern of dependence appears in the autocorrelation function of the fitted residuals. Figure 19 shows the 137 year limit cycle of Model 9 with its ascent and descent periods. The limit cycle is asymmetric with a range in amplitude of 17.7 to 94.5 and an average ascent/descent period of 4.3/6.23 years versus 4.6/6.58 years for the actual yearly sunspot numbers from 1700 to 1920 (Table 1). In comparing Model 9's limit cycle (Figure 19) with the real yearly sunspot data (Figure 13) note that the standard deviation of the fitted residual's error variance is estimated as $(MSS)^{1/2} = 10.69$ sunspots.

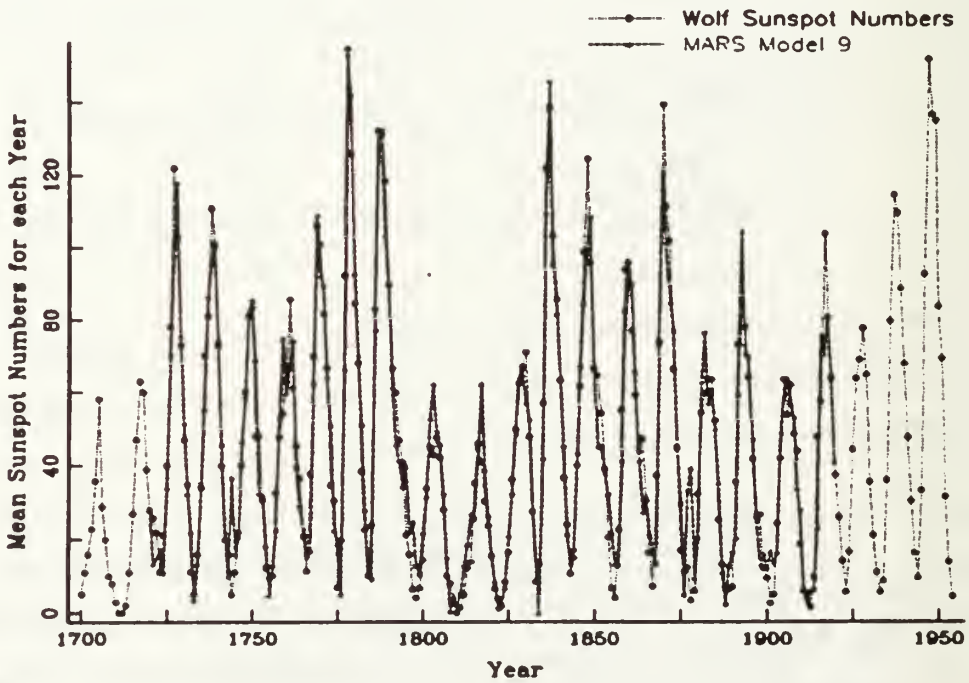

Figure 13. The yearly Wolf sunspot numbers (1700-1955) versus the fit of ASTAR Model 9 (1720-1920). The yearly sunspot numbers (1700-1719) were used for initialization. The yearly sunspot numbers (1921-1955) were used to examine the prediction performance of ASTAR Model 9 and other models of the yearly sunspot numbers.

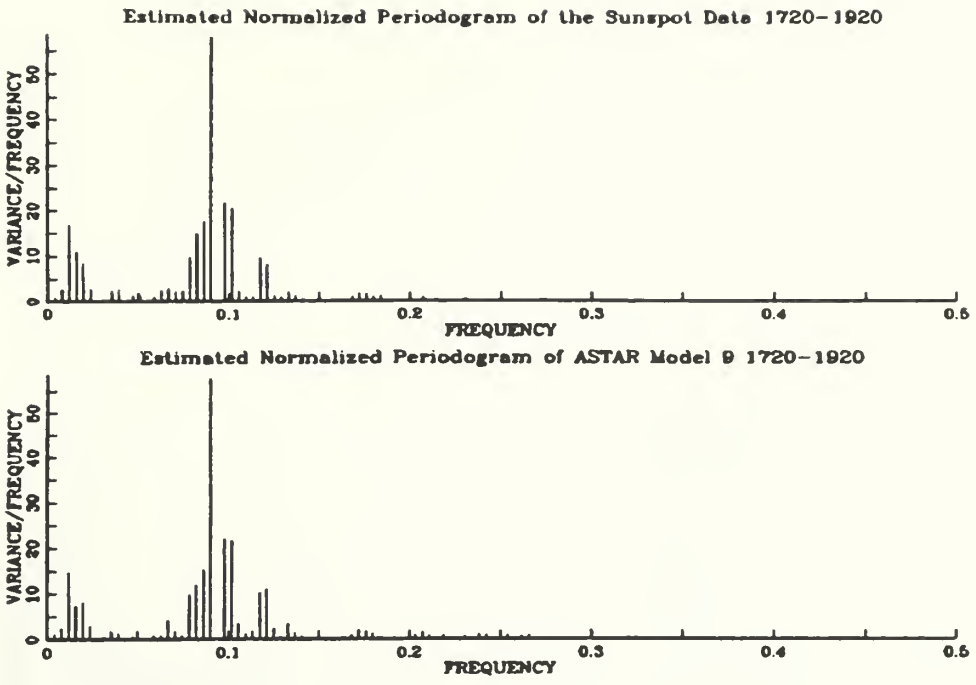

Figure 14. The estimated normalized periodogram of the yearly Wolf sunspot numbers (1720-1920) [top] versus the estimated normalized periodogram of ASTAR Model 9 (1720-1920) [bottom]. The broad conclusion from the top periodogram is that there is a rather diffuse cycle in the data with a period of about 11 years, and a longer period of about 67 years.

Figure 20 is a graphical representation of ASTAR Model 9. Each column in the plot represents an individual term in equation (38) that is identified along the plot's horizontal axis, e.g., (1) represents the $X_{\tau-1}$ term (second term of line one in equation (38)) while (1),(2) represents the 2-way interaction term between $X_{\tau-1}$ and $X_{\tau-2}$ (first term of line two in equation (38)). The vertical axis defines the range of values of the yearly sunspot numbers during the modeling period from 1720-1920. The plot lines (1-way interaction), symbols (2-way interaction) and the combination of lines and symbols (3-way interaction) define the range of yearly sunspot number values that permit a nonzero contribution to the value of $\widehat{X}_\tau$ by a term of the model. Located underneath the plot is summary information of the contributions by each model term during the modeling period to include the number (Num) of times each model term made a nonzero contribution to the value of $\widehat{X}_\tau$ and

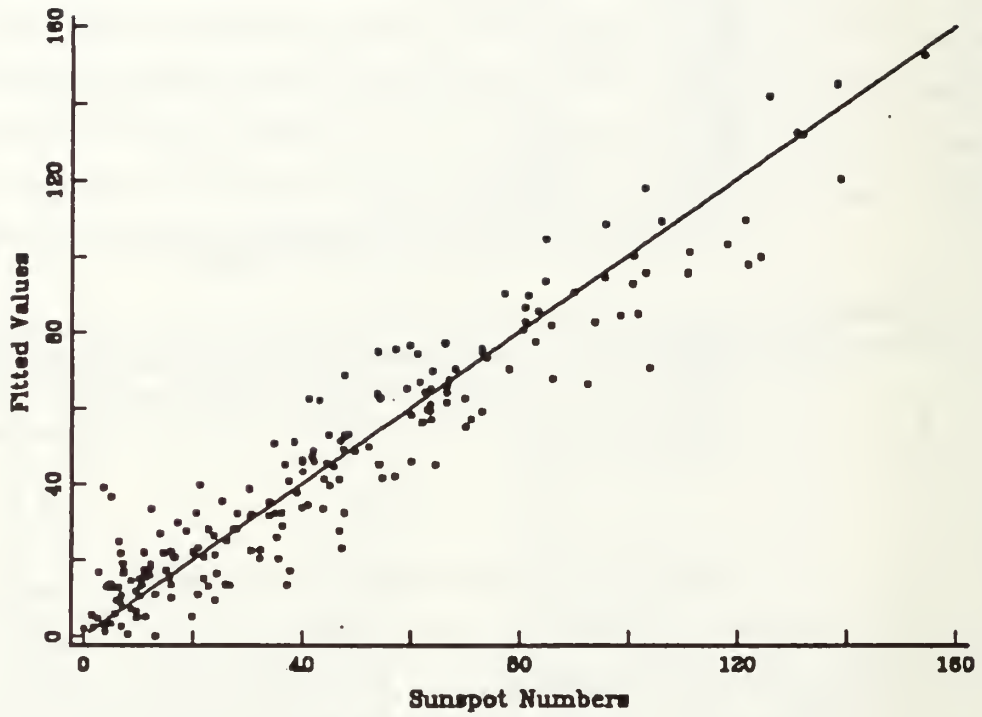

Figure 15. The empirical quantile-quantile plot for the fitted values of ASTAR Model 9 versus the yearly Wolf sunspot numbers for the period 1720-1920. No obvious pattern exists.

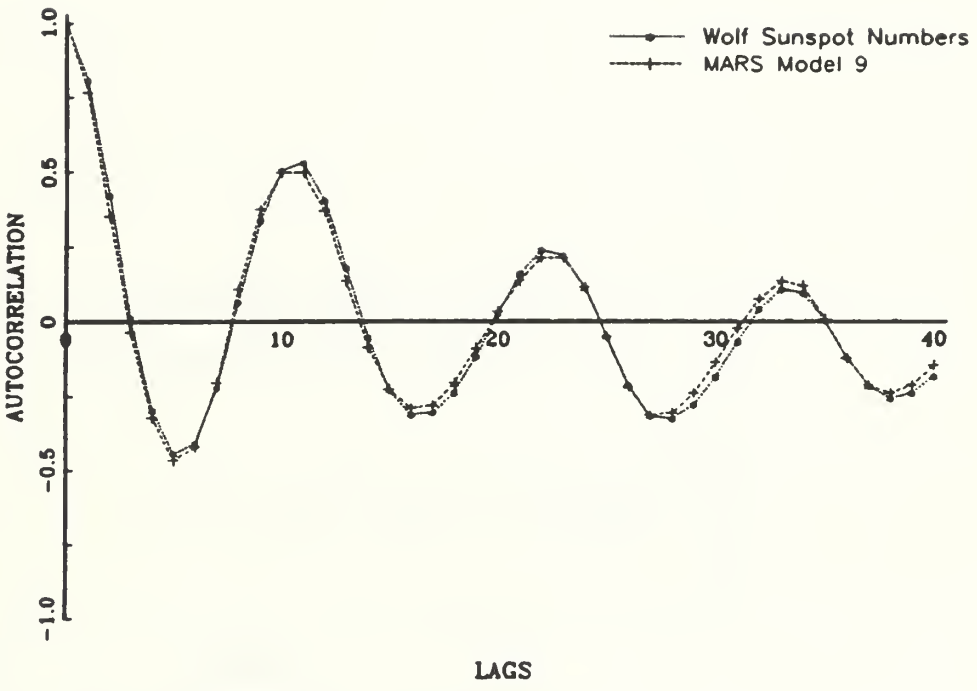

Figure 16. The autocorrelation functions of the yearly Wolf sunspot numbers and ASTAR Model 9 for the period 1720-1920. The dominant cycle of period approximately 11 years is clearly evident.

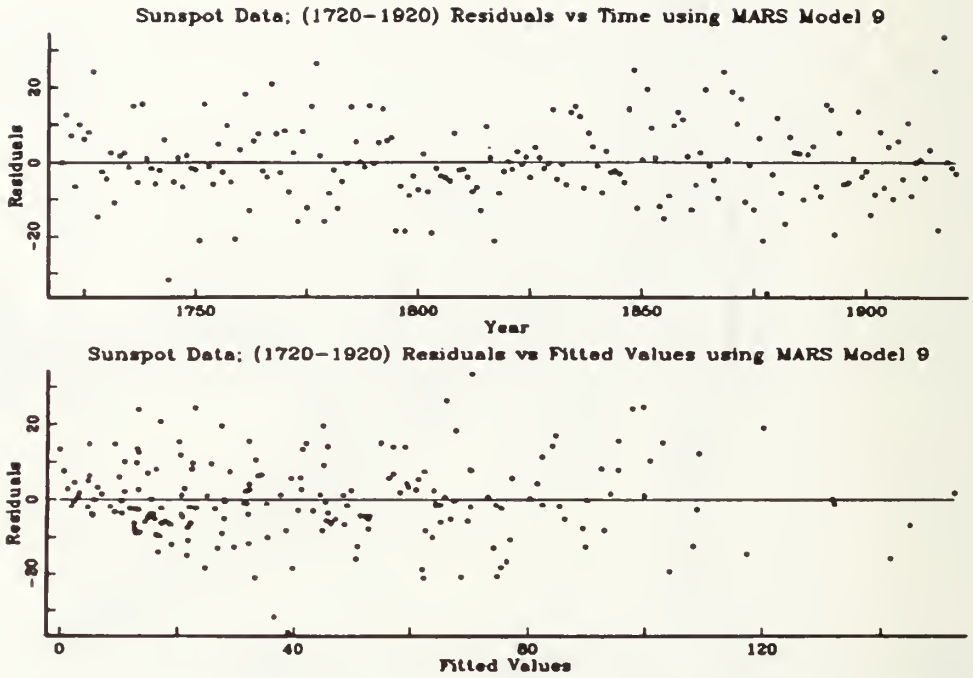

Figure 17. Fitted residuals from ASTAR Model 9 of the yearly Wolf sunspot numbers (1720-1920) versus year [top]. Fitted residuals versus the fitted yearly sunspot numbers from ASTAR Model 9 of the yearly Wolf sunspot numbers (1720-1920) [bottom].

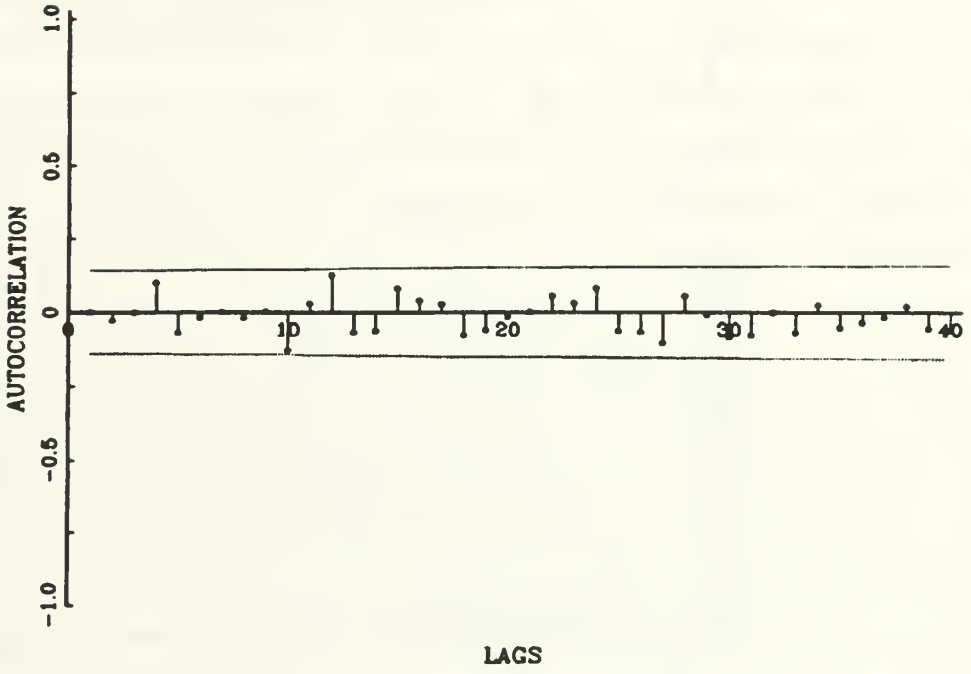

Figure 18. The autocorrelation function (first 40 lags) of the fitted residuals for ASTAR Model 9 of the yearly Wolf sunspot numbers (1720-1920). There is no pattern of dependence in the residuals. The confidence bounds are approximate, individual confidence bands.

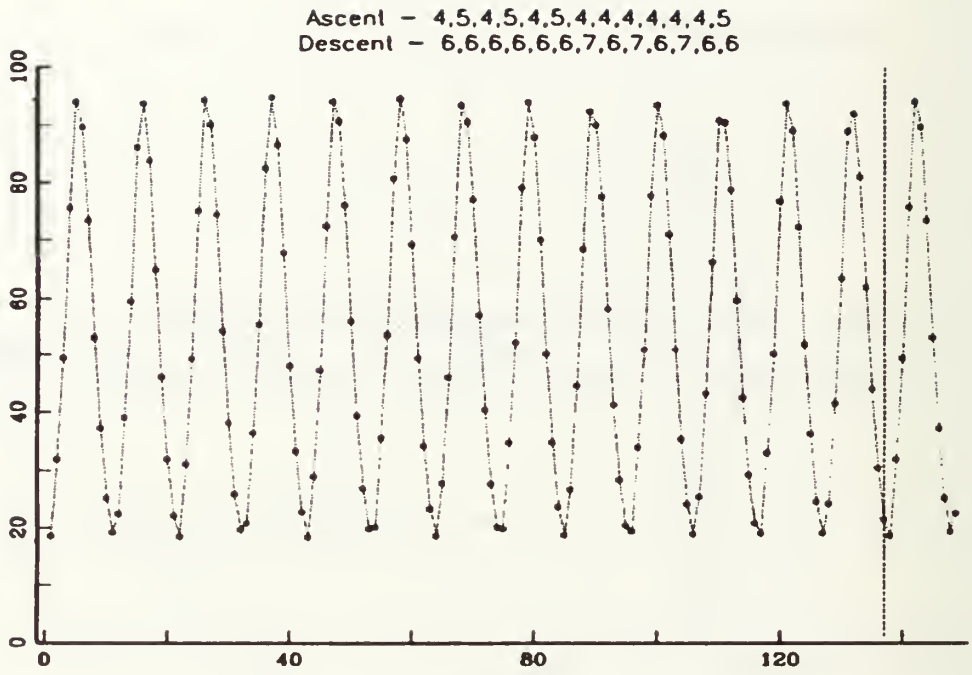

Figure 19. The limit cycle for ASTAR Model 9 of the yearly Wolf sunspot numbers (1720-1920). The limit cycle is 137 years long with the indicated ascent and descent periods. The limit cycle is generated using ASTAR Model 9 initialized with the yearly sunspot numbers (1700-1719). The 'subcycles' have lengths of 10 or 11 years with 4 or 5 years per ascent period and 6 or 7 years per descent period.

the minimum (Min), mean (Mean) and maximum (Max) values of each term's nonzero contributions.

The key point for discussing a graphical representation for a model such as that given in Figure 20 is that it can be used to analyze the use for and contribution of each of the terms in an ASTAR model. For example, using the ASTAR model of the yearly sunspot numbers given in equation (38) and graphically displayed in Figure 20, the 2-way interaction term (1),(4), which represents the term $-.032X_{\tau-1}(17.1 - X_{\tau-4})_+$, has a nonzero contribution to the value of $\hat{X}_\tau$ if and only if $X_{\tau-1} > 0$ and $X_{\tau-4} < 17.1$. Using Figures 20 and 13 it can be seen that this term's maximum contribution arises when $X_{\tau-1}$ is at a yearly sunspot cycle peak and $X_{\tau-4}$ is in a yearly sunspot cycle trough. Thus the purpose of this term is to initiate the downward turn of $\hat{X}_\tau$ to the next yearly sunspot cycle trough. Another example is the 3-way interaction term (1),(2),(5), which represents the term $.004X_{\tau-1}(26.0 - X_{\tau-2})_+(X_{\tau-5} - 41.0)_+$. This term has a nonnegative contribution to the value of $\hat{X}_\tau$ if and only if $X_{\tau-1} > 0$, $X_{\tau-2} < 26$ and $X_{\tau-5} > 41$. Again using Figures 20 and 13 it can be seen that this term's maximum contribution arises when $X_{\tau-1}$ and $X_{\tau-2}$ are in a yearly sunspot cycle trough and $X_{\tau-5}$ is at a yearly sunspot cycle peak. Thus the purpose of this term is to initiate the upward turn of $\hat{X}_\tau$ to the next yearly sunspot cycle peak. Likewise, the (1),(2) term has a large contribution when both $X_{\tau-1}$ and $X_{\tau-2}$ are high. Like the (1),(4) term, the (1),(2) term is used to initiate the downward turn of $\hat{X}_\tau$ to the next yearly sunspot cycle trough. However, unlike the (1),(4) term whose number of nonzero contributions to the value of $\hat{X}_\tau$ is severely restricted due to the threshold at 17.1, the (1),(2) term continues to drive $\hat{X}_\tau$ into the next trough until $X_{\tau-2} \leq 26$. Similar analysis can be performed on other model terms or combinations of model terms.

Other useful graphical displays for the ASTAR models of the yearly sunspot numbers are the individual plots of each term's contribution to the value of $\hat{X}_\tau$ versus sunspot number year. These plots complement Figure 20 and permit the comparison of the magnitude and location of each term's contribution. In summary, the graphical displays mentioned above provide a valuable analytical tool for studying nonlinear time series models such as those developed with the ASTAR methodology.

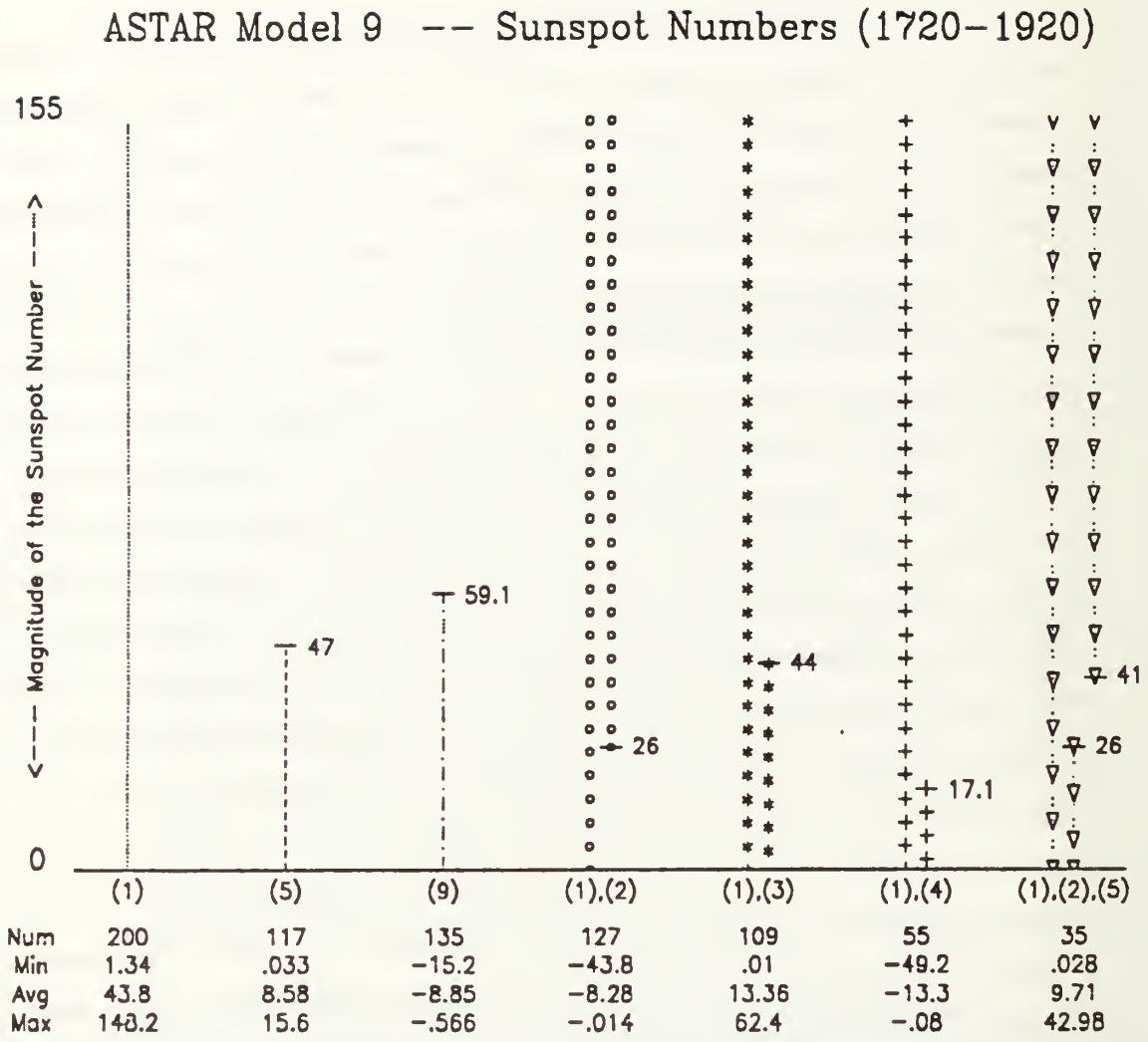

Figure 20. Graphical representation of ASTAR Model 9 given in Equation (38) of the yearly Wolf sunspot numbers (1720-1920). Each column in the plot represents a term of the model whose contributions to the value of $\widehat{X}_t$ is summarized underneath the plot. Lines in the first three columns, labeled (1) and (5) and (9) define the range of values for nonzero contributions to the value of $\widehat{X}_t$ by the *linear* terms X_{t-1}, X_{t-5} or X_{t-9} respectively; symbols in the next three columns, labeled (1),(2) and (1),(3) and (1),(4), define the range of values for nonzero contributions to the value of $\widehat{X}_t$ by the *2-way interaction* terms $X_{t-1}X_{t-2}$, $X_{t-1}X_{t-3}$ and $X_{t-1}X_{t-4}$ respectively; and in the last column the combination of lines and symbols define the range of values for nonzero contributions to the value of $\widehat{X}_t$ by the *3-way interaction* term $X_{t-1}X_{t-2}X_{t-5}$.

c. *Sunspot Number Prediction using ASTAR Models*

The predictive performance of ASTAR Model 9 (38) was investigated by comparing its forward-step predictions with the forward-step predictions of other models that were developed using the 221 yearly sunspot numbers from 1700-1920. These include forward-step predictions (Moeanaddin, 1989) for the 35 yearly sunspot numbers from 1921-1955 using the Full Linear Autoregressive, Bilinear Subset (Rao, 1984) and Self Exciting Threshold, SETAR, (Tong, 1983) models. The Full Linear Autoregressive (39) and Bilinear Subset (40) models used the 10 sunspots from 1700-1709 for initialization while the SETAR (41) and ASTAR (38) models used the 20 sunspots from 1700-1719 for initialization.

Full Linear Autoregressive

$$\hat{X}_\tau = \begin{cases} 32.55 & +1.216X_{\tau-1} - 0.467X_{\tau-2} - 0.142X_{\tau-3} \\ & +0.169X_{\tau-4} - 0.147X_{\tau-5} + 0.054X_{\tau-6} \\ & -0.053X_{\tau-7} + 0.067X_{\tau-8} + 0.113X_{\tau-9} \end{cases} \tag{39}$$

Bilinear Subset

$$\hat{X}_\tau = \begin{cases} 6.886 & +1.501X_{\tau-1} - 0.767X_{\tau-2} + 0.115X_{\tau-9} - 0.014X_{\tau-2}\epsilon_{\tau-1} \\ & +0.006X_{\tau-8}\epsilon_{\tau-1} - 0.007X_{\tau-1}\epsilon_{\tau-3} + 0.006X_{\tau-4}\epsilon_{\tau-3} \\ & +0.004X_{\tau-1}\epsilon_{\tau-6} + 0.004X_{\tau-2}\epsilon_{\tau-4} + 0.002X_{\tau-3}\epsilon_{\tau-2} \end{cases} \tag{40}$$

Self Exciting Threshold (SETAR)

$$\hat{X}_\tau = \begin{cases} 10.544 & +1.692X_{\tau-1} - 1.159X_{\tau-2} + 0.236X_{\tau-3} + 0.150X_{\tau-4} \\ & \text{if } X_{\tau-3} \leq 36.6 \\ 7.804 & +0.743X_{\tau-1} - 0.041X_{\tau-2} - 0.202X_{\tau-3} + 0.173X_{\tau-4} \\ & -0.227X_{\tau-5} + 0.019X_{\tau-6} + 0.161X_{\tau-7} - 0.256X_{\tau-8} \\ & +0.319X_{\tau-9} - 0.389X_{\tau-10} + 0.431X_{\tau-11} - 0.397X_{\tau-12} \\ & \text{if } X_{\tau-3} > 36.6 \end{cases} \tag{41}$$

The forward-step predictions for each of these models are obtained by fixing both the model terms and coefficients during the entire prediction period. The mean sum of squares for the errors of the predictions (PMSE) obtained by these models and ASTAR Model 9 are given in Table 4.

TABLE 4. FORWARD-STEP PREDICTIONS OF THE YEARLY WOLF SUNSPOT NUMBERS: The mean sum of squares error $\hat{\sigma}_\epsilon^2$, number of model parameters and the predictive mean sum of squares error $\hat{\sigma}_\epsilon^2(i)$ for the ith forward-step prediction for the period (1921-1955) of the Full Linear Autoregressive (AR), Bilinear Subset, SETAR and ASTAR models of the yearly Wolf sunspot numbers for the period (1700-1920).

Model	AR (Full)	Bilinear (Subset) (Rao)	SETAR (Tong)	ASTAR Model 9
$\hat{\sigma}_\epsilon^2$	199.3	124.3	153.7	114.1
Number of Parameters	10	11	19	14
$\hat{\sigma}_\epsilon^2(1)$	190.9	123.8	153.9	132.5
$\hat{\sigma}_\epsilon^2(2)$	414.8	337.6	388.4	314.8
$\hat{\sigma}_\epsilon^2(3)$	652.1	569.7	672.7	467.3
$\hat{\sigma}_\epsilon^2(4)$	797.3	621.3	641.2	415.1
$\hat{\sigma}_\epsilon^2(5)$	770.8	718.4	835.3	367.2
$\hat{\sigma}_\epsilon^2(6)$	786.4	732.4	900.7	408.0
$\hat{\sigma}_\epsilon^2(7)$	789.0	781.7	993.8	441.2
$\hat{\sigma}_\epsilon^2(8)$	827.8	833.2	1083.6	455.2

The performance of the ASTAR model for forecasting the yearly sunspot numbers from 1921-1955 is a considerable improvement over the AR and Threshold models for every forward step, and it is an improvement over the Bilinear Subset model for every forward step except the first step. Also, it is interesting and surprising to note that the predictive mean sum of squares error for the ASTAR model decreases in the fourth and fifth step before increasing again. This phenomenon was also identified in subsequent analysis

of other ASTAR models that have limit cycles. We attribute this interesting phenomenon to the underlying limit cycle of the models (Tong, 1985, and Moeanaddin, 1989).

While the prediction of the yearly sunspot numbers for 1921-1955 is a considerable improvement over the previous threshold and bilinear modeling efforts, it may be difficult to justify using the kth forward-step prediction as a conditional expectation when making the ith forward-step prediction of an ASTAR model with a threshold on $X_{\tau-j}$ and $i > k \geq j$. Tong (1983, 1985) suggests, as one approach to this problem, 'moving the trigger', i.e., prohibit a threshold from forming on a lagged variable with lag less than the desired maximum forward-step prediction. Tong (1983) reported several TAR models that 'moved the trigger' and were used for prediction of the sunspot numbers. TAR Model AS7133 (42) was developed with a threshold value on $X_{\tau-7}$ using the sunspot numbers from 1700-1890 and used to obtain the forward-step predictions of the sunspot numbers from 1921-1955 (Moeanaddin, 1989).

TAR Model AS7133

$$\hat{X}_\tau = \begin{cases} 9.267 + 0.987X_{\tau-1} - 0.307X_{\tau-2} - 0.108X_{\tau-3} + 0.166X_{\tau-4} \\ \quad - 0.297X_{\tau-5} + 0.285X_{\tau-6} - 0.155X_{\tau-7} - 0.171X_{\tau-8} \\ \quad + 0.210X_{\tau-9} - 0.041X_{\tau-10} + 0.353X_{\tau-11} - 0.196X_{\tau-12} \\ \qquad\qquad \text{if} \quad X_{\tau-7} \leq 58.55 \\ \\ 26.159 + 1.577X_{\tau-1} - 1.240X_{\tau-2} \\ \qquad\qquad \text{if} \quad X_{\tau-7} > 58.55 \end{cases} \tag{42}$$

To incorporate this idea, MARS was used to formulate several models of the yearly sunspot numbers by 'moving the trigger'. This is simple to do since in the input to MARS one can specify that the predictor variables are not permitted to have a knot i.e., are linear if included. For this modeling effort the interest was to permit prediction of approximately one sunspot number cycle. Thus the lagged variables with lag less than or equal to eleven were not permitted to form knots. Note that the modeling effort was

restricted to 1700-1890 to correspond to modeling efforts by Tong (1983) that also 'moved the trigger'.

Table 5 gives the forward-step predictions from 2 ASTAR models created using MARS with the restriction that thresholds were prohibited on $X_{\tau-1}$ thru $X_{\tau-11}$. These two new models both have a single threshold on $X_{\tau-14}$ and thus permit up to a 14 step ahead (full yearly sunspot cycle) forecast of the yearly sunspot numbers without the difficulties discussed in the previous paragraph. Model GCV9-322 (43) has 8 coefficients and includes a 4-way interaction term while Model GCV9-1028 (44) has 11 coefficients and includes only linear and 2-way interaction terms.

ASTAR Model GCV9-322

$$\hat{X}_\tau = \begin{cases} 10.760 & +1.326X_{\tau-1} - 0.714X_{\tau-2} \\ & -0.003X_{\tau-1}X_{\tau-12} + 0.568X_{\tau-2}X_{\tau-11} \\ & +.0002X_{\tau-1}X_{\tau-6}(X_{\tau-14} - 73.9)_+ \\ & -.000003X_{\tau-1}X_{\tau-2}X_{\tau-6}(X_{\tau-14} - 73.9)_+ \end{cases} \tag{43}$$

ASTAR Model GCV9-1028

$$\hat{X}_\tau = \begin{cases} -11.256 & +1.257X_{\tau-1} + 0.576X_{\tau-9} \\ & -0.008X_{\tau-1}X_{\tau-2} + 0.002X_{\tau-2}X_{\tau-3} - 0.002X_{\tau-2}X_{\tau-5} \\ & +0.003X_{\tau-2}X_{\tau-11} - 0.006X_{\tau-3}X_{\tau-9} - 0.003X_{\tau-9}X_{\tau-10} \\ & +0.004X_{\tau-1}(X_{\tau-14} - 60.0)_+ \end{cases} \tag{44}$$

As with the previous predictions of the yearly sunspot numbers from 1921-1955 with ASTAR Model 9 (Table 4), the Bilinear Subset model has the best MSS for the first forward-step prediction. However, from the second forward step for GCV9-322 (third forward step for GCV9-1028), the ASTAR models have the best predictive mean sum of squares error and are again a considerable improvement over the Full Linear Autoregressive, Bilinear Subset and SETAR models for the period 1921-1955. Also, again note that the

TABLE 5. FORWARD-STEP PREDICTIONS OF THE YEARLY WOLF SUNSPOT NUMBERS: The mean sum of squares error $\hat{\sigma}_\epsilon^2$, number of model parameters and the predictive mean sum of squares error $\hat{\sigma}_\epsilon^2(i)$ for the ith forward-step prediction for the period (1921-1955) of the AR, Bilinear Subset, SETAR and ASTAR models of the yearly Wolf sunspot numbers. *Here, in contrast to the ASTAR model used for the results in Table 4, thresholds were not permitted for lagged predictor variables in MARS unless the lag was greater than eleven.* The modeling period for the AR and Bilinear Subset models is (1700-1920) while the modeling period for the SETAR and ASTAR models is (1700-1890).

Model	AR	Bilinear Subset (Rao)	SETAR AS7133 (Tong)	ASTAR Model GCV9-322	ASTAR Model GCV9-1028
$\hat{\sigma}_\epsilon^2$	199.3	124.3	152.3	155.5	149.9
Number of Parameters	10	11	17	8	11
$\hat{\sigma}_\epsilon^2(1)$	190.9	123.8	161.9	158.3	205.1
$\hat{\sigma}_\epsilon^2(2)$	414.8	337.6	362.6	333.4	425.0
$\hat{\sigma}_\epsilon^2(3)$	652.1	569.7	593.2	515.3	472.2
$\hat{\sigma}_\epsilon^2(4)$	797.3	621.3	650.1	449.1	416.6
$\hat{\sigma}_\epsilon^2(5)$	770.8	718.4	613.2	404.2	402.6
$\hat{\sigma}_\epsilon^2(6)$	786.4	732.4	584.8	377.7	384.0
$\hat{\sigma}_\epsilon^2(7)$	789.0	781.7	508.1	373.4	378.6
$\hat{\sigma}_\epsilon^2(8)$	827.8	833.2	531.8	372.8	391.0
$\hat{\sigma}_\epsilon^2(9)$	862.1	900.6	518.8	319.1	389.8
$\hat{\sigma}_\epsilon^2(10)$	895.6	961.9	520.9	302.7	379.5
$\hat{\sigma}_\epsilon^2(11)$	982.9	1013.8	563.0	297.3	371.0
$\hat{\sigma}_\epsilon^2(12)$	1168.5	1139.2	650.5	361.9	419.1

predictive mean sum of squares error for the ASTAR and SETAR models decrease after several forward steps before increasing again.

Moeanaddin (1989) used the AR, Bilinear and SETAR models for 'risky' prediction for the roughly 2 yearly sunspot cycle period from 1956-1979. This prediction period is 'risky' because it includes an 'abnormal' jump in the yearly sunspot numbers from 38.0 in 1955 to 141.7 in 1956. The forward-step PMSE's of the SETAR models are slightly better than the MSS's of the ASTAR models for this period, although the potential of the ASTAR models developed in this chapter were not fully explored. However, the bilinear model's predictive performance is rather explosive. Moeanaddin (1989) indicates that the collapse of the bilinear models prediction may be due to its non-invertibility and the effect of the influential observation in 1956.

G. SUMMARY

MARS is a new nonparametric regression modeling methodology, due to Friedman, that utilizes low-order regression spline modeling and a modified recursive partitioning strategy to exploit the localized low-dimensional behavior of the data used to construct $\hat{f}(\boldsymbol{x})$. Although MARS is a computationally intensive regression methodology, it provides a systematic methodology for deriving nonlinear threshold models for high-dimensional data that are naturally continuous in the domain of the predictor variables, and can have multiple partitions and predictor variable interactions.

By letting the predictor variables in MARS be lagged values of a time series, one obtains an adaptive spline threshold autoregressive (ASTAR) model, which is a new method for nonlinear modeling of time series that extends the threshold autoregression methodology developed by Tong (1985). A significant feature of ASTAR when modeling time series data with periodic behavior is its ability to produce continuous models for the regression function with underlying sustained oscillations (limit cycles). An initial analysis of the yearly Wolf sunspot numbers (1700-1890) and (1700-1920) using ASTAR produced several models with underlying limit cycles. When used to predict the yearly sunspot numbers (1921-1955), the ASTAR models are a significant improvement over existing Threshold and Bilinear models.

An important aspect of any overall regression modeling effort is the interpretation and analysis that answers questions about the model's behavior and reveals relationships

between the response variable (output) and predictor variables (input). However, the functional form of an ASTAR model, with its combination of different predictor variables and multiple threshold values, makes its straightforward interpretation and analysis difficult. In this regard a graphical representation was developed to permit the interpretation and analysis of ASTAR Model 9 of the Wolf sunspot numbers. Further enhancements are obtained by integrating color in the graphical representation. It was shown that this graphical representation can be used to analyze the use for and contribution of each of the terms in an ASTAR model.

III. SEMI-MULTIVARIATE NONLINEAR MODELING OF TIME SERIES SYSTEMS USING MULTIVARIATE ADAPTIVE REGRESSION SPLINES (MARS)

A. INTRODUCTION

While ASTAR models of univariate time series certainly have widespread applicability, the identification of semi-multivariate threshold autoregressive models that consider the complex interactions within a time series system would have even greater applicability. In this chapter the ASTAR methodology developed in Chapter II is extended to the *semi-multivariate* ASTAR modeling of a time series system. This builds upon semi-multivariate threshold autoregressive (TAR) modeling by Tong et al. (1985). Thus MARS is used to model a single response variable of a time series system using predictor variables that are the lagged values of both the response and input time series. For example, for $\tau = 1, 2, \ldots N$, let $\{Y_\tau\}$ and $\{Z_\tau\}$ be time series that represent system inputs and $\{X_\tau\}$ be a times series representing the system output. The set of possible predictor variables for this semi-multivariate time series system are $X_{\tau-1}, \ldots, X_{\tau-d_1}; Y_\tau, \ldots, Y_{\tau-d_2}$ and $Z_\tau, \ldots, Z_{\tau-d_3}$, where the maximum lags d_1, d_2 and d_3 are not necessarily equivalent. Also, $d_1 + (d_2 + 1) + (d_3 + 1) = p$, the total number of predictor variables. If MARS is applied to this system of predictor variables the result is a semi-multivariate ASTAR model that seems well suited for taking into account the complex interactions among the multivariate, cross-correlated, lagged predictor variables of a time series system. The analysis of an Icelandic river using past riverflow, temperature and precipitation to develop semi-multivariate ASTAR models extends earlier TAR modeling of this Icelandic riverflow. Note that the same problem for normal multivariate linear time series processes such as ARMA models may be treated by Kalman filtering (see, e.g., Gelb, 1974). However, here we are not concerned with complete multivariate models, in the sense of Box and Tiao (1977) and Tiao and Tsay (1989).

B. SEMI-MULTIVARIATE NONLINEAR TIME SERIES MODELING USING MARS

There are numerous semi-multivariate times series systems that appear well suited for analysis using the MARS methodology, such as sea surface temperatures using lagged temperature, surface winds and time as predictor variables (Breaker and Lewis, 1985; Altman, 1987) or riverflow using lagged river flow, temperature and precipitation as predictor variables (Gudmundsson, 1970; Tong et al., 1985). One possible source of nonlinearity in the riverflow system might occur due to the change in temperature above and below freezing. Below freezing, precipitation (snow) does not 'runoff' as rapidly as precipitation (rain) at higher temperatures. Other applications exist for any multivariate times series system with suspected nonlinear behavior, if the objective is to model a single output stream given multiple input streams to the system. In particular in Chapter IV, a series of sea surface temperatures will be analyzed. What is of more interest, as noted above, is to model the current sea surface temperatures as a function of lagged sea surface temperatures, lagged wind shear (wind velocity squared) and lagged wind direction.

To provide a framework for the semi-multivariate time series model, suppose that for $\tau = 1,2,\ldots,N$, $\{Y_\tau\}$ and $\{Z_\tau\}$ denote the input time series and $\{X_\tau\}$ the output time series for a time series system we wish to model. The complete description for the general form of a semi-multivariate time series model is very complex. However, using the notation $\|$ (from Tong, 1985) to separate the possible predictor variables of each different time series and following (1), we can nominally describe X_τ with the semi-multivariate time series regression model

$$X_\tau = f(1 \,\|\, X_{\tau-1}, X_{\tau-2}, \ldots, X_{\tau-d_1} \,\|\, Y_\tau, Y_{\tau-1}, \ldots, Y_{\tau-d_2} \,\|\, Z_\tau, Z_{\tau-1}, \ldots, Z_{\tau-d_3}) + \epsilon_\tau, \quad (45)$$

where $f(\cdot)$ represents some functional form of its argument, 1 denotes a model constant, and the maximum lags d_1, d_2, and d_3 are not necessarily equivalent. Also, Y_τ and Z_τ, the current values of the predictive time series, may or may not be included in (45), depending on the time series system and the use for which the model is to be put. Generally, prediction of X_τ at τ would preferably be done without the knowledge of Y_τ and Z_τ. This is because if X_τ is measurable, it will generally be known only when Y_τ and Z_τ are finally known.

Both Tong (1985) and Tsay (1989) suggest a methodology for semi-multivariate TAR modeling that follows the TAR methodology for a univariate time series, i.e., identification of linear semi-multivariate autoregressive time series models in each disjoint subregion of the predictor variable space. For example, their notation for a *very simple* two-subregion semi-multivariate TAR model based on a single partition in the space of all the predictor variables at, say, $Z_\tau = 3$ is

$$\hat{X}_\tau = \begin{cases} (0.5 \parallel 1.1 \parallel -2.7, 1.1 \parallel 4.3, -2.8) & \text{if } Z_\tau \leq 3 \\ \\ (2.3 \parallel 0.1, -0.2 \parallel 1.7 \parallel -0.1, 2.1) & \text{if } Z_\tau > 3, \end{cases}$$

which represents the model

$$\hat{X}_\tau = \begin{cases} 0.5 + 1.1X_{\tau-1} - 2.7Y_\tau + 1.1Y_{\tau-1} + 4.3Z_\tau - 2.8Z_{\tau-1} & \text{if } Z_\tau \leq 3 \\ \\ 2.3 + 0.1X_{\tau-1} - 0.2X_{\tau-2} + 1.7Y_\tau - 0.1Z_\tau + 2.1Z_{\tau-1} & \text{if } Z_\tau > 3. \end{cases} \tag{46}$$

The semi-multivariate TAR methodologies of Tong (1985) and Tsay (1989) focus on univariate and bivariate scatterplot analysis and on the evaluation of empirical percentiles of preselected threshold variable candidates. *These methods are also permitted with* MARS. However, the predictor variables of a time series system may possess physical behavior not readily apparent when we restrict our modeling methodology to the above approach. The key point is that Tong's and Tsay's methods are time consuming, generally limited to one or two dimensions and may not be sufficient for identifying changes in the physical behavior of a nonlinear time series system. Thus, a semi-multivariate TAR model is still burdened with the limitations of a univariate TAR model, i.e., a threshold model created with the piecewise linear models from each disjoint subregion of a domain D of the predictor variables. Also the TAR model is usually discontinuous at each subregion boundary (threshold) and is limited to a small number of thresholds, most often using only one variable, due to the difficulties associated with the threshold selection process.

The MARS methodology supplements Tong's (1985) and Tsay's (1989) approach by admitting a more general class of continuous nonlinear semi-multivariate threshold models than permitted with the semi-multivariate TAR methodology, and by providing a more

systematic (automatic) way of fitting the model. The methodology for developing this class of nonlinear semi-multivariate threshold models is called SMASTAR (Semi-Multivariate Adaptive Spline Threshold Autoregression). Following Chapter II, the fact that one obtains from the MARS algorithm a more *general* class of continuous nonlinear semi-multivariate threshold models than permitted with semi-multivariate TAR methodology (for example a model such as (46)) can be shown using a simple example.

Let X_τ be a time series we wish to model with the lagged predictor variables $X_{\tau-1}$, $X_{\tau-2}, Y_{\tau-1}, Y_{\tau-2}, Z_{\tau-1}$ and $Z_{\tau-2}$. Also, let the notation $(U-t)_+^\pm$ represent $(t-U)_+$ and $(U-t)_+$ where $(u)_+ = u$ if $u \geq 0$ and 0 otherwise. Extending the example for the ASTAR time series model developed in Chapter II, each forward step of the MARS algorithm selects *one and only one* set of new terms for the SMASTAR time series model from the candidates specified by previously selected terms of the model. For our example problem the sets of candidates in the *initial* forward step of the MARS algorithm are

$$\begin{aligned}
&(X_{\tau-1} - t_x^*)_+^\pm \quad \text{or} \quad (X_{\tau-2} - t_x^*)_+^\pm \quad \text{or} \\
&(Y_{\tau-1} - t_y^*)_+^\pm \quad \text{or} \quad (Y_{\tau-2} - t_y^*)_+^\pm \quad \text{or} \\
&(Z_{\tau-1} - t_z^*)_+^\pm \quad \text{or} \quad (Z_{\tau-2} - t_z^*)_+^\pm,
\end{aligned} \tag{47}$$

where t_x^*, t_y^* and t_z^* are unknown partition points (thresholds) in the range of their respective lagged predictor variable. For our example problem, assume that the MARS algorithm selects the lagged predictor variable $X_{\tau-2}$ with threshold value $t_x^* = t_1$, i.e., $(X_{\tau-2} - t_1)_+$ and $(t_1 - X_{\tau-2})_+$ are the initial terms (other than the constant) in the SMASTAR time series model. The sets of candidates in the second forward step of the MARS algorithm includes *all univariate candidates in* (47) and the new sets of multivariate candidates (interactions):

$$\begin{aligned}
&(X_{\tau-1} - t_x^*)_+^\pm (X_{\tau-2} - t_1)_+, \quad \text{or} \quad (X_{\tau-1} - t_x^*)_+^\pm (t_1 - X_{\tau-2})_+, \quad \text{or} \\
&(Y_{\tau-1} - t_y^*)_+^\pm (X_{\tau-2} - t_1)_+, \quad \text{or} \quad (Y_{\tau-1} - t_y^*)_+^\pm (t_1 - X_{\tau-2})_+, \quad \text{or} \\
&(Y_{\tau-2} - t_y^*)_+^\pm (X_{\tau-2} - t_1)_+, \quad \text{or} \quad (Y_{\tau-2} - t_y^*)_+^\pm (t_1 - X_{\tau-2})_+, \quad \text{or} \\
&(Z_{\tau-1} - t_z^*)_+^\pm (X_{\tau-2} - t_1)_+, \quad \text{or} \quad (Z_{\tau-1} - t_z^*)_+^\pm (t_1 - X_{\tau-2})_+, \quad \text{or} \\
&(Z_{\tau-2} - t_z^*)_+^\pm (X_{\tau-2} - t_1)_+, \quad \text{or} \quad (Z_{\tau-2} - t_z^*)_+^\pm (t_1 - X_{\tau-2})_+,
\end{aligned} \tag{48}$$

due to the initial selection of $(X_{\tau-2}-t_1)_+$ and $(t_1-X_{\tau-2})_+$ as terms in the SMASTAR time series model. One and only one of the terms from (47) or (48) is selected for inclusion in the model in the next forward step of the MARS algorithm. It follows that SMASTAR time series models could have multiple thresholds on one variable, say $X_{\tau-2}$ in our example, by again selecting $(X_{\tau-2}-t_x^*)_+^{\pm}$ in (47) for some new partition point $t_x^* \neq t_1$. The forward-step algorithm continues at each step by selecting the set of univariate or multivariate terms that, for a given threshold t_x^*, t_y^* or t_z^* discovered using exhaustive search, most contributes to "improving" model fit. The sets of candidates for each subsequent forward step of the SMASTAR algorithm is nondecreasing in size and is based on previously selected terms of the model. As discussed in Chapter II the forward-step algorithm is followed by a backward-step algorithm that trims excess terms of the model that no longer sufficiently contribute to the model fit. And again, both the forward and backward steps of the algorithm use GCV^* (23) to evaluate model fit versus model complexity (Chapter V discusses alternative model selection criteria).

Again, for $\tau = 1,2,\ldots,N$, let $\{Y_\tau\}$ and $\{Z_\tau\}$ denote the input time series and $\{X_\tau\}$ the output time series for a time series system that we wish to model. Let the p predictor variables in MARS for the τth value in a time series $\{X_\tau\}$ be; $X_{\tau-1}, X_{\tau-2}, \ldots, X_{\tau-d_1}, Y_\tau, Y_{\tau-1}, \ldots, Y_{\tau-d_2}$, and $Z_\tau, Z_{\tau-1}, \ldots, Z_{\tau-d_3}$, which we represent as $\boldsymbol{X}_{\tau-1}^{d_1}$, $\boldsymbol{Y}_\tau^{d_2+1}$, and $\boldsymbol{Z}_\tau^{d_3+1}$, respectively. Following (30), the functional form of the SMASTAR model that estimates X_τ is

$$\hat{X}_\tau = \sum_{j=1}^{S} c_j K_j(\boldsymbol{X}_{\tau-1}^{d_1}, \boldsymbol{Y}_\tau^{d_2+1}, \boldsymbol{Z}_\tau^{d_3+1}) \tag{49}$$

so that $\hat{X}_\tau$ is an additive function of the product spline basis functions $\left\{K_j(\boldsymbol{X}_{\tau-1}^{d_1}\boldsymbol{Y}_\tau^{d_2+1}, \boldsymbol{Z}_\tau^{d_3+1})\right\}_{j=1}^{S}$ associated with the subregions $\{R_j\}_{j=1}^{S}$. As with the AS-TAR time series model (31), the functional form of the SMASTAR time series model may be expanded using the ordered sequences of truncated spline functions (20 and 21) that define each product spline basis function.

Let a and b be dummy variables that index the ordered sequence of truncated spline functions $\left\{T_{a,\boldsymbol{r}_b}(\boldsymbol{X}_{\tau-1}^{d_1}\boldsymbol{Y}_\tau^{d_2+1}, \boldsymbol{Z}_\tau^{d_3+1})\right\}_{j=1}^{S}$ such that $0 \leq a < b \leq j$. Also to account for the additional complexity of a multivariate time series system let $\boldsymbol{r}_j = (\pm v, t, l)$ represent a 3-

tuple associated with the truncated spline function $T_{a,r_b}(\boldsymbol{X}_{\tau-1}^{p})$ whose components identify: $\pm$, the direction of the truncated spline (left or right); v, the specific predictor variable; t, the partition point; and l, the input time series used as predictor variables. Given this additional notation the functional form of the SMASTAR time series model for the τth value in a time series $\{X_\tau\}$ using this expansion is

$$\hat{X}_\tau = \sum_{j=1}^{S} c_j \prod_{T_{a,r_b} \in K_j} [sgn_v(l_{\tau-v} - t)]_+ \tag{50}$$

where the argument $\boldsymbol{X}_{\tau-1}^{d_1}, \boldsymbol{Y}_{\tau}^{d_2+1}, \boldsymbol{Z}_{\tau}^{d_3+1}$ of $T_{a,r_b}(\boldsymbol{X}_{\tau-1}^{d_1}, \boldsymbol{Y}_{\tau}^{d_2+1}, \boldsymbol{Z}_{\tau}^{d_3+1})$ and $K_j(\boldsymbol{X}_{\tau-1}^{d_1}, \boldsymbol{Y}_{\tau}^{d_2+1}, \boldsymbol{Z}_{\tau}^{d_3+1})$ is suppressed for simplicity. Again, note that the truncated spline functions act in only one dimension although their argument is a vector of predictor variables.

By modeling a time series system using the MARS algorithm, we overcome some of the limitations of the semi-multivariate TAR modeling approach. The MARS methodology provides a systematic procedure for deriving a nonlinear semi-multivariate time series model that is naturally continuous in the domain of the predictor variables. As shown in Chapter II and later in Chapter IV with the yearly Wolf sunspot numbers and Granite Canyon data sets, ASTAR models of univariate time series can possess multiple thresholds and high level predictor variable interactions. This construction has now been extended to the multivariate setting with SMASTAR models, which can also possess multiple thresholds and high level predictor variable interactions. However, now the threshold values and predictor variable interactions can take place among the cross-correlated, lagged predictor variables of a semi-multivariate time series system. In contrast, the semi-multivariate TAR methodologies of Tong (1985) and Tsay (1989) focus on scatterplot analysis and the evaluation of empirical percentiles of preselected threshold variable candidates, which is time consuming and may not be sufficient for identifying changes in the physical behavior of a nonlinear time series system. Also the semi-multivariate TAR model is still burdened with the limitations of a univariate TAR model, i.e., a discontinuous threshold model created with the piecewise linear models from several disjoint subregions of a domain D of the predictor variables.

1. Semi-Multivariate Non Linear Threshold Modeling of the Vatnsdalsa River

As an illustration of SMASTAR's ability to model an actual semi-multivariate times series system, the riverflow (Tong, 1985) for the Vatnsdalsa River in Iceland from 1972 to 1974 is analyzed. A riverflow at a given location and time is an output of a complex time series system with inputs that include aspects of the geography, geology, meteorology and topography within the river's region of flow. Extensive literature is available on the modeling complexities of riverflow and will not be revisited here other than to state that the use and control of riverflow is of great concern in many countries of the world. Also, *riverflow data generally has a very non-normal distribution*, in part because of the nonlinear seasonal variations of the system variables and in part because of the difficulty in capturing all of the influential variables within the modeling effort. Lawrance and Kottegoda (1977) provide an excellent historical review of statistical hydrology and discuss stochastic modeling of riverflow with the goal that " ... models should be able to reproduce, in simulation, sequences of flows or lake levels or rainfalls, which are statistically indistinguishable from the relevant historical sequence." This prescription in essence permits prediction and the study of physical changes that can affect the hydrological system, e.g., a dam in the case of a riverflow system.

The Vatnsdalsa riverflow data, Figure 21, consists of the river's average rate of daily flow (X_t) in $m^3/sec.$, the daily precipitation (Y_t) in mm, and the average daily temperature (Z_t) in ^oC, at the Hveravellir meteorological station in Iceland for the period from 1972 to 1974. The range of values for daily riverflow for this period is 3.67 to 54.0 $m^3/sec.$, with a mean value of 8.94 $m^3/sec.$; the range of values for daily precipitation for this period is 0.0 to 79.3 mm, with a mean of 2.51 mm; the range of values for daily temperature for this period is -22.4 to 13.9 ^oC with a mean value of $-.44^oC$. Both the riverflow and temperature are highly autocorrelated times series with lag 1 correlations of .92 and .90 respectively. The precipitation record is actually translated forward by one day due to the difference in the time during the day for recording the precipitation data and the time during the day for recording the temperature and riverflow data. An extensive discussion of the Vatnsdalsa riverflow system is provided by Gudmundsson (1970) and Tong et al. (1985).

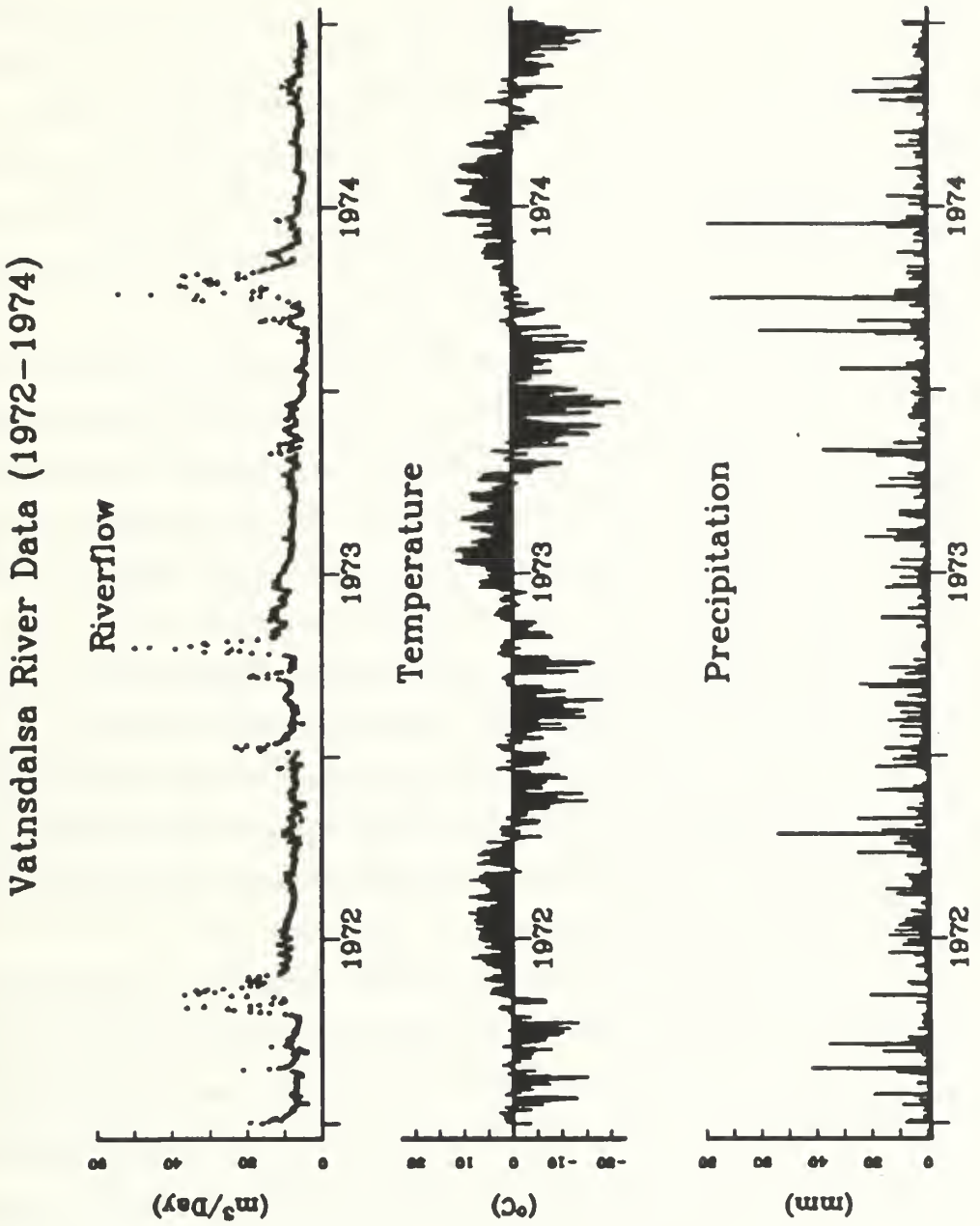

Figure 21. The record of daily Vatnsdalsa riverflow, temperature and precipitation for 1972 to 1974 taken at the Hveravellir meteorological station in Iceland for the period from 1972 to 1974. The range of values for daily riverflow for this period is 3.67 to 54.0 $m^3/sec.$, with a mean value of 8.94 $m^3/sec.$; the range of values for daily precipitation for this period is 0.0 to 79.3 mm, with a mean of 2.51 mm; the range of values for daily temperature for this period is -22.4 to 13.9 $°C$ with a mean value of $-.44°C$. The precipitation record is actually translated forward by one day due to the difference in the time during the day for recording the precipitation data and the time during the day for recording the temperature and riverflow data.

Although the length of the riverflow data is relatively short, 1096 days, it does provide a basis for SMASTAR model development and analysis. The primary interest for modeling the Vatnsdalsa riverflow with MARS was to determine if MARS could produce reasonable SMASTAR models in a semi-multivariate time series setting, and if so, could the SMASTAR model identify reasonable nonlinear structure in the riverflow data, e.g., changes to riverflow due to temperatures above and below freezing? Finally, could the SMASTAR model be used for prediction of riverflow one day ahead?

Graphical analysis of the riverflow data, Figure 21, reveals an extremely high riverflow that occurs each spring. Precipitation that falls in the form of snowfall during the winter accumulates until the temperature rises sufficiently in the spring to release it to the riverflow system. Note that the high riverflow corresponds to the general rise in temperature during the early months of each year. Also, a considerable shift in the overall riverflow occurs in 1974 that is not evident during the previous two years. The spring riverflow during 1974 is higher and of longer duration than the spring riverflows for the previous two years. This surge gives way to the extremely low riverflow in the latter half of 1974, that again is not characteristic of the same period riverflow for the previous two years. These severe changes in the riverflow structure for 1974 can be attributed to a combination of extremely high rainfall and the rapid warming of the snowpack that occurred earlier in 1974 than in the previous two years.

The empirical density functions for the riverflow [top], temperature [middle] and precipitation [bottom] data are shown in Figure 22. This figure should be interpreted with the understanding that the data is clearly seasonal. The y-axis scale (density) of the plots are equal while the x-axis is scaled for each time series and reflects the range of each time series for the years 1972-1974. The empirical density function of the temperature is relatively symmetric while the empirical density function of the riverflow and temperature data are extremely skewed. The skewness in the precipitation data is a result of the heavy but infrequent precipitation that occurs each year. The skewness in the riverflow data can be attributed to the high riverflow that occurs each spring. The skewed distributions of these data sets suggest the possible use of transformations for symmetry (normality) and variance stabilization to moderate the influence of the extreme values. Transformations were considered for the precipitation data. However, riverflow is the output stream that we

are modeling. Therefore, we chose to deal directly with this data to avoid the difficulties associated with inverse transformations for purposes of analysis and prediction.

Many other modeling methodologies could be used for modeling this type of semi-multivariate hydrological data. One method, previously discussed, is to develop semi-multivariate TAR models for various regions of the predictor variable space. The semi-multivariate TAR modeling effort of the Vatnsdalsa River data is briefly discussed in the next section. Other methods consider models using a fixed signal with noise. However, as with the yearly Wolf sunspot numbers, attempts to model the data with a fixed cycle period signal plus (possibly correlated) noise have failed because the cyclical component in the spectrum for this riverflow system is quite spread out. In particular, using Figure 21, note the size and shift in time of the Vatnsdalsa's riverflow that takes place in the spring of 1974, as compared to the Vatnsdalsa's riverflow during the spring for the previous two years.

a. *TAR and SMASTAR Models of the Vatnsdalsa River (1972-1974)*

Tong et al. (1985) considered a series of semi-multivariate linear and TAR time series models for the Vatnsdalsa riverflow data (1972-1974). Their goal, to develop nonlinear models for purposes of simulation along with establishing relationships between riverflow and important meteorological variables met with limited success due to the limitations of the TAR methodology. Also, the TAR models included Y_τ and Z_τ, i.e., same day precipitation and temperature. Although a model that includes Y_τ and Z_τ does permit analysis of the "immediate" influence of temperature and precipitation on riverflow, it also essentially prohibits the use of the model for riverflow prediction. Several semi-multivariate time series models from Tong et al. (1985) are of interest.

The first model from Tong et al. (1985), Tong Model 1 of the Vatnsdalsa river system, shown at Figure 23, is the ordinary semi-multivariate *linear* time series model for riverflow during 1972 with only precipitation and temperature as the system inputs, i.e., without lagged riverflow as a model predictor variable. Tong Model 1 is

$$X_\tau = 9.40 + 0.17Y_\tau + 0.11Y_{\tau-1} - 0.07Z_\tau + \epsilon_\tau, \tag{51}$$

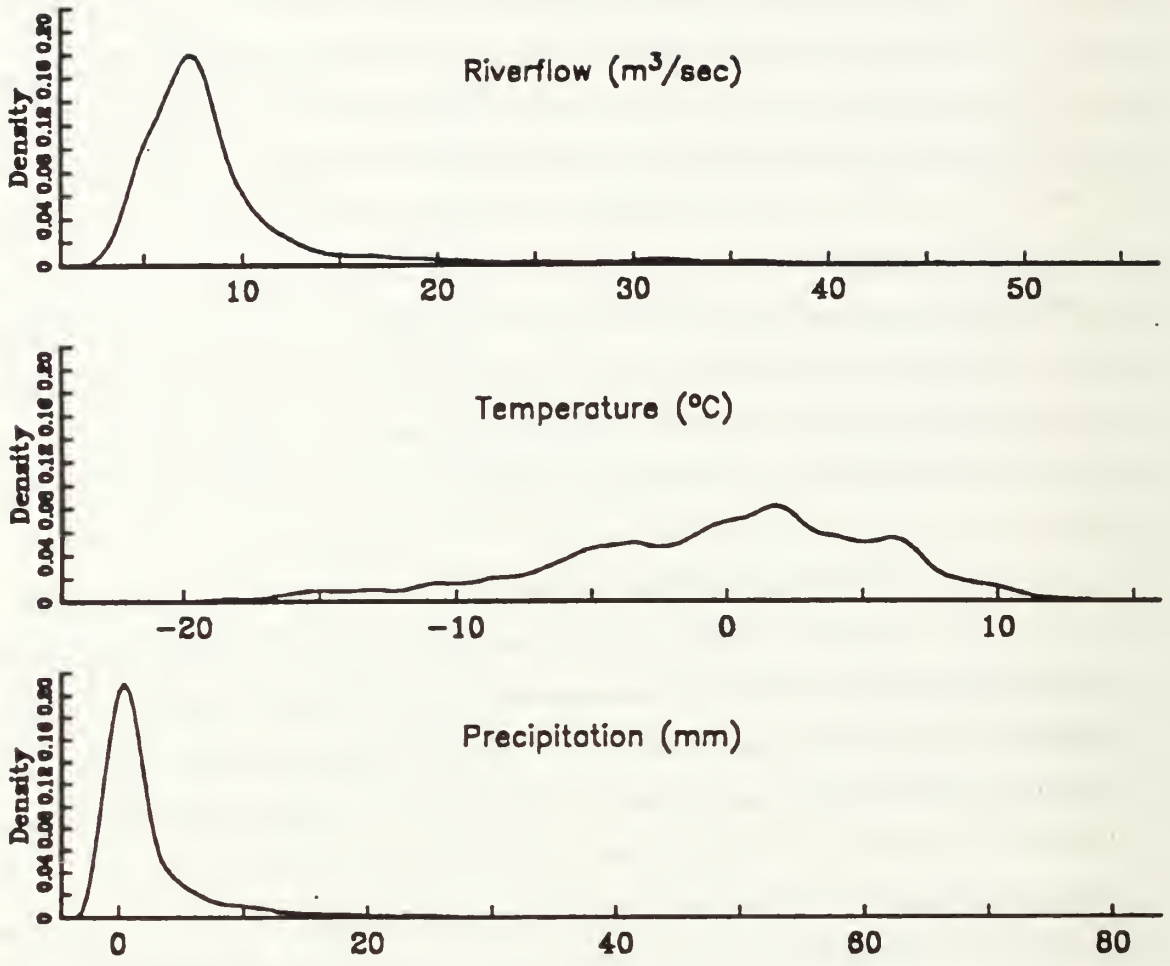

Figure 22. The empirical density functions of the riverflow, temperature and rainfall data for the Vatnsdalsa riverflow system for 1972 to 1974 taken at the Hveravellir meteorological station in Iceland. The y-axis scale (density) of the plots are equal while the x-axis is scaled for each time series and reflects the range of each time series for the years 1972-1974. The empirical density function of the temperature is relatively symmetric. In contrast, the empirical density functions of the riverflow and temperature data are extremely skewed. The skewness in the precipitation data is a result of the heavy but infrequent precipitation that occurs each year, while the skewness in the riverflow data can be attributed to the high riverflow that occurs each spring. This figure should be interpreted with the understanding that the data is clearly seasonal. The

where ϵ is assumed to be Gaussian white noise and the standard error of the fitted residuals $\sigma_\epsilon = 4.64$. The poor quality of this *linear* model is revealed by its inability to capture the sharp structure of high riverflow during the spring runoff. Also from Tong et al. (1985), the magnitude of σ_ϵ is larger than the average value of X_τ and it is difficult to explain the rational for the negative coefficient on the precipitation variable Z_τ. The shortcomings of this model indicate the importance of lagged riverflow to help capture the structure of the riverflow system.

The second model from Tong et al. (1985), Tong Model 2 of the Vatnsdalsa river system, shown at Figure 24, is the ordinary semi-multivariate *linear* time series model for riverflow during 1972 with precipitation, temperature and riverflow as the system inputs. Tong Model 2 is

$$\begin{aligned} X_\tau = \; .73 \quad & + 1.12X_{\tau-1} - 0.23X_{\tau-2} + 0.12X_{\tau-3} - 0.09X_{\tau-4} \\ & + 0.09Y_\tau - 0.03Y_{\tau-1} - 0.04Y_{\tau-2} \\ & + 0.01Z_\tau + 0.07Z_{\tau-1} - 0.06Z_{\tau-2} + 0.02Z_{\tau-3} + \epsilon_\tau, \end{aligned} \tag{52}$$

where ϵ is assumed to be Gaussian white noise and the standard error of the fitted residuals $\sigma_\epsilon = 1.68$. To simplify the presentation of more complex semi-multivariate models that follow, using notation from Tong (1985), Tong Model 2 may be rewritten as

$$\begin{aligned} X_t = \; (.73 \quad & \| \, 1.12, -0.23, 0.12, -0.09 \\ & \| \, 0.09, -0.03, -0.04 \\ & \| \, 0.01, 0.07, -0.06, 0.02) \quad + \epsilon_\tau \end{aligned}$$

where $\|$ is used to separate the coefficients of the lagged predictor variables from the different time series. The fitted values and residuals of Tong Model 2, shown in Figure 24, are a considerable improvement over those for Tong Model 1, shown in Figure 23. The analysis of Tong Model 2, using equation (52), reveals the immediate and lagged influence of all three different predictor variables. Also, in the absence of present and lagged rainfall and assuming that present and lagged temperatures are close to $0°C$, i.e., $Y_\tau, Y_{\tau-1}$ and $Y_{\tau-2}$, and $Z_\tau, Z_{\tau-1}, Z_{\tau-2}$, and $Z_{\tau-3} = 0$, this model's riverflow reaches a steady state flow of about

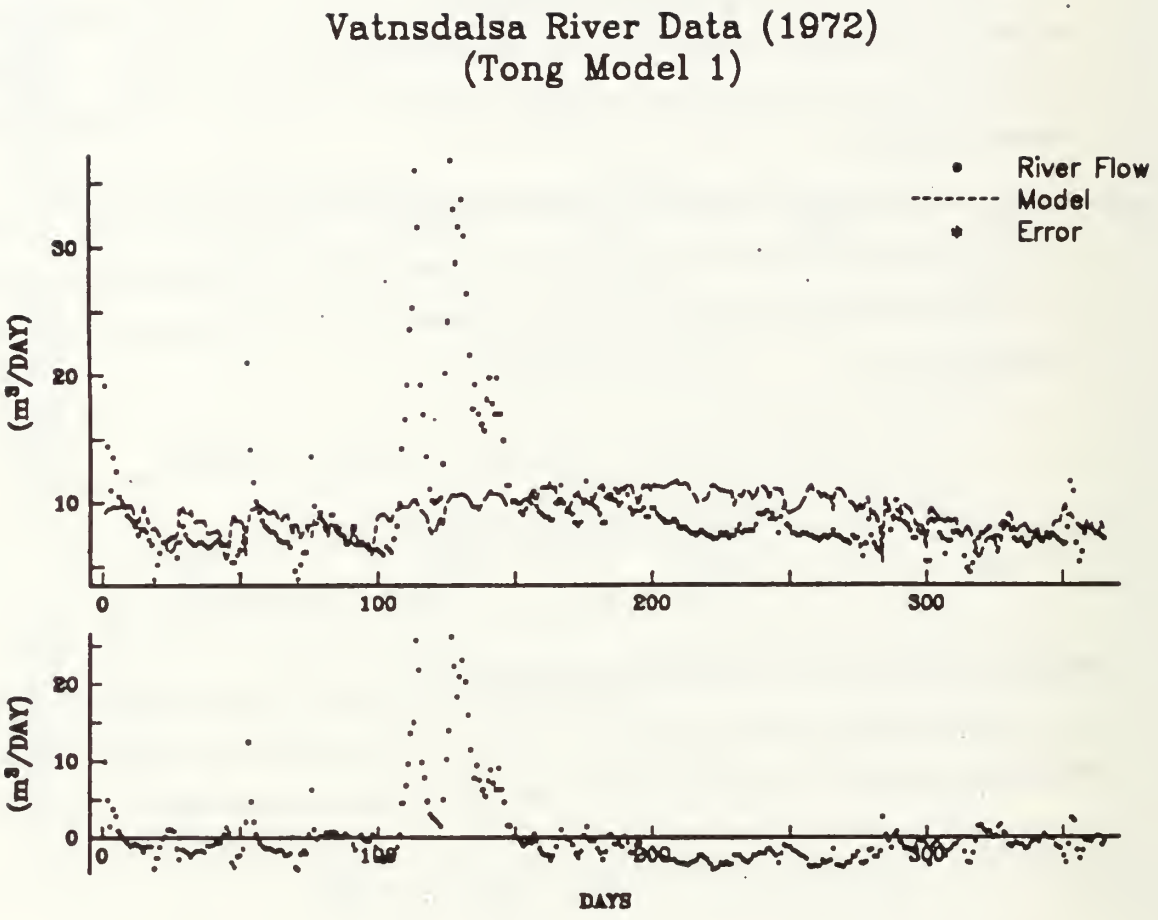

Figure 23. Vatnsdalsa riverflow data during 1972 versus the fitted (predicted) values (top) and residuals (bottom) for the ordinary semi-multivariate *linear* time series model, Tong Model 1, from Tong et al. (1985). Tong Model 1 uses precipitation $Y_{\tau-i}$, and temperature $Z_{\tau-i}$ as system inputs i.e., lagged riverflow $X_{\tau-i}$ is not used as a predictor variable. The standard error of the fitted residuals, $\sigma_\epsilon = 4.64$, is larger than the average value of the riverflow. The poor quality of this model's fit is revealed by its inability to capture the sharp structure of high riverflow during the spring runoff. This model reveals the importance of lagged riverflow to help capture the structure of the riverflow system.

$9m^3$/sec., which is reasonable based on the analysis of the riverflow data in Figure 21. However, there are still several instances when the fitted values of Model 2 deviate from the structure of the actual riverflow (the fitted values actually indicate negative riverflow). These deviations occur most often during the period when the temperature is rapidly rising during early spring and thus is indicative of the nonlinear relationships that exist among the predictor variables in this time series system.

In response to the difficulties of the ordinary semi-multivariate linear time series models, Tong et al. (1985) proposes several semi-multivariate TAR models for the Vatnsdalsa riverflow system. The methodology for developing a semi-multivariate TAR models was discussed at the beginning of Section B. The progressive use of this methodology resulted in a final semi-multivariate TAR model from Tong et al. (1985), Tong Model 5 of the Vatnsdalsa river system, using lagged riverflow $X_{\tau-1},\ldots,X_{\tau-10}$, lagged temperature $Y_\tau,\ldots,Y_{\tau-10}$, and lagged precipitation $Z_\tau,\ldots,Z_{\tau-10}$, as the predictor variables. Using the notation developed from (46) and (52), Tong Model 5 for the period 1972 to 1974 is,

$$
\begin{aligned}
X_t &= \begin{cases} (0.75 & \| \ 1.06,-0.26,0.09,-0.11,0.08 \\ & \| \ 0.02,-0.03,0.01,-0.02 \\ & \| \ -0.02,-0.01,-0.00,0.01) + \epsilon_\tau^1 \end{cases} && \text{if } Z_\tau \le -2, \\
&= \begin{cases} (1.21 & \| \ 0.97,-0.29,0.04,0.11 \\ & \| \ 0.53,0.02,-0.02 \\ & \| \ 0.03,0.12,-0.04,-0.02) + \epsilon_\tau^2 \end{cases} && \text{if } -2 < Z_\tau \le 2, \\
&= \begin{cases} (1.97 & \| \ 1.38,-0.70,0.47,0.02,-0.19,-0.02,0.34,-0.23 \\ & \| \ -0.59,0.07,-0.11,-0.05,0.07,0.13,-0.25 \\ & \| \ 0.03,0.03,-0.01,0.04,-0.03,-0.04,0.13,0.01) + \epsilon_\tau^3 \end{cases} && \text{if } 2 < Z_\tau \le 5, \\
&= \begin{cases} (0.59 & \| \ 1.22,-0.49,0.30,-0.17,0.27,-0.26,0.11 \\ & \| \ -0.02,-0.02,0.01,-0.01,0.02,-0.04 \\ & \| \ 0.01,0.01,-0.01,-0.01,0.01,-0.02) + \epsilon_\tau^4 \end{cases} && \text{if } Z_\tau > 5, \qquad (53)
\end{aligned}
$$

where each regions errors, $\{\epsilon_\tau^i\}_{i=1}^4$, are assumed to be Gaussian white noise sequences that are independent of each other and where the standard error of the *pooled* fitted residuals

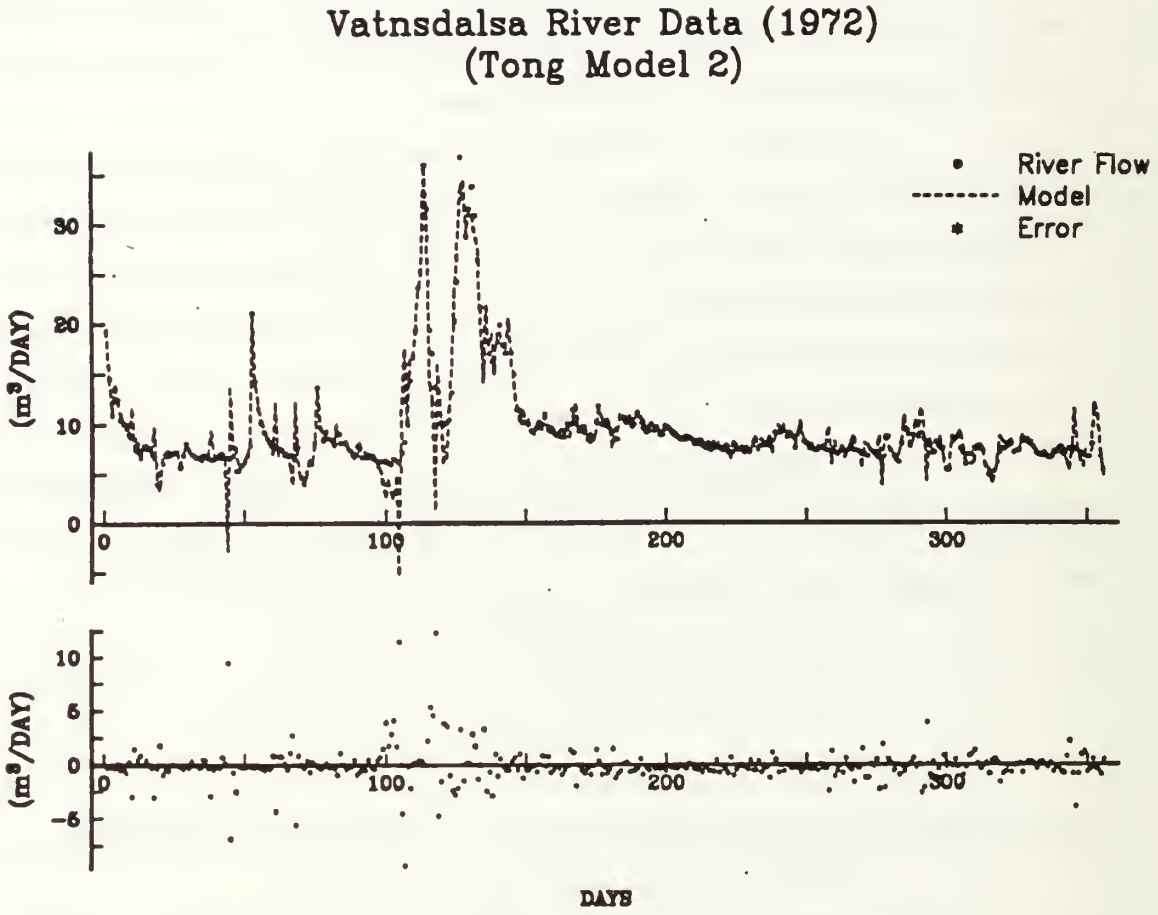

Figure 24. Vatnsdalsa riverflow data during 1972 versus the fitted (predicted) values (top) and residuals (bottom) for the ordinary semi-multivariate *linear* time series model, Tong Model 2, from Tong et al. (1985), using precipitation $Y_{\tau-i}$, temperature $Z_{\tau-i}$ and riverflow $X_{\tau-i}$ as system inputs. The standard error of the fitted residuals $\sigma_\epsilon = 1.68$. The fit of this model is a considerable improvement over Tong Model 1, Figure 23. However, there are still several instances when the fitted values of the model deviate from the structure of the actual riverflow (the fitted values actually indicate negative riverflow). This occurs most often during the period when the temperature is rapidly rising during early spring and is indicative of the nonlinear relationships that exist among the predictor variables in this time series system.

is $\sigma_\epsilon = 1.69 m^3/sec$. Tong Model 5 uses 73 *parameters* in four disjoint subregions of the predictor variables that are described by the univariate thresholds, $-2, 2$ and $5\,^oC$, on the *lag 0 temperature variable* Z_τ.

Figure 25 is a plot of the fitted values (top) and residuals (bottom) for Tong Model 5 of the Vatnsdalsa River system. Tong Model 5 appears to capture the overall structure of the riverflow data within each disjoint subregion described by the model. Figure 26 contains plots of the autocorrelation [top] and normalized cumulative periodogram [bottom] of the fitted residuals from the second subregion of Tong Model 5 ($-2 < Z_\tau < 2$). If the fitted residuals of Tong Model 5 are truly independent then the fitted residuals in each of the four model subregions should also be independent. The fitted residual autocorrelation plot, with approximate individual 95% confidence intervals for zero correlation, shows that significant short term residual autocorrelation still exists. Also, the normalized cumulative periodogram plot, with a reference line for the normalized spectrum of Gaussian white noise and 90% Kolmogorov-Smirnov (K-S) bounds, shows that we should reject the hypothesis that the fitted residuals from the second subregion are Gaussian white noise. Note that these results are similar for the other subregions of the model.

In summary, although Tong Model 5 appears to capture the overall structure of the riverflow data, this and other semi-multivariate TAR models of the Vatnsdalsa riverflow system were unable to model the data in such a way as to produce riverflow data with Gaussian or even uncorrelated residuals. This in conjunction with the enormous size of this semi-multivariate TAR model (73 parameters) may reflect the inability of the TAR methodology to capture the complex predictor variable interactions present in this riverflow system. Note that the maximum lag of a time series used for predictor variables in a semi-multivariate TAR model, e.g., equation (53), may be different from subregion to subregion. However, within a subregion the semi-multivariate TAR model is of full size, i.e., all autoregressive terms for each input time series up to the maximum lag are included. Thus there is no subset selection of the predictor variables in the semi-multivariate TAR model. In contrast, the SMASTAR methodology permits subset selection of the lagged predictor variables used from each input time series.

Given the preliminary analysis of the Vatnsdalsa riverflow system and the semi-multivariate TAR modeling effort, the MARS algorithm was used to develop SMAS-

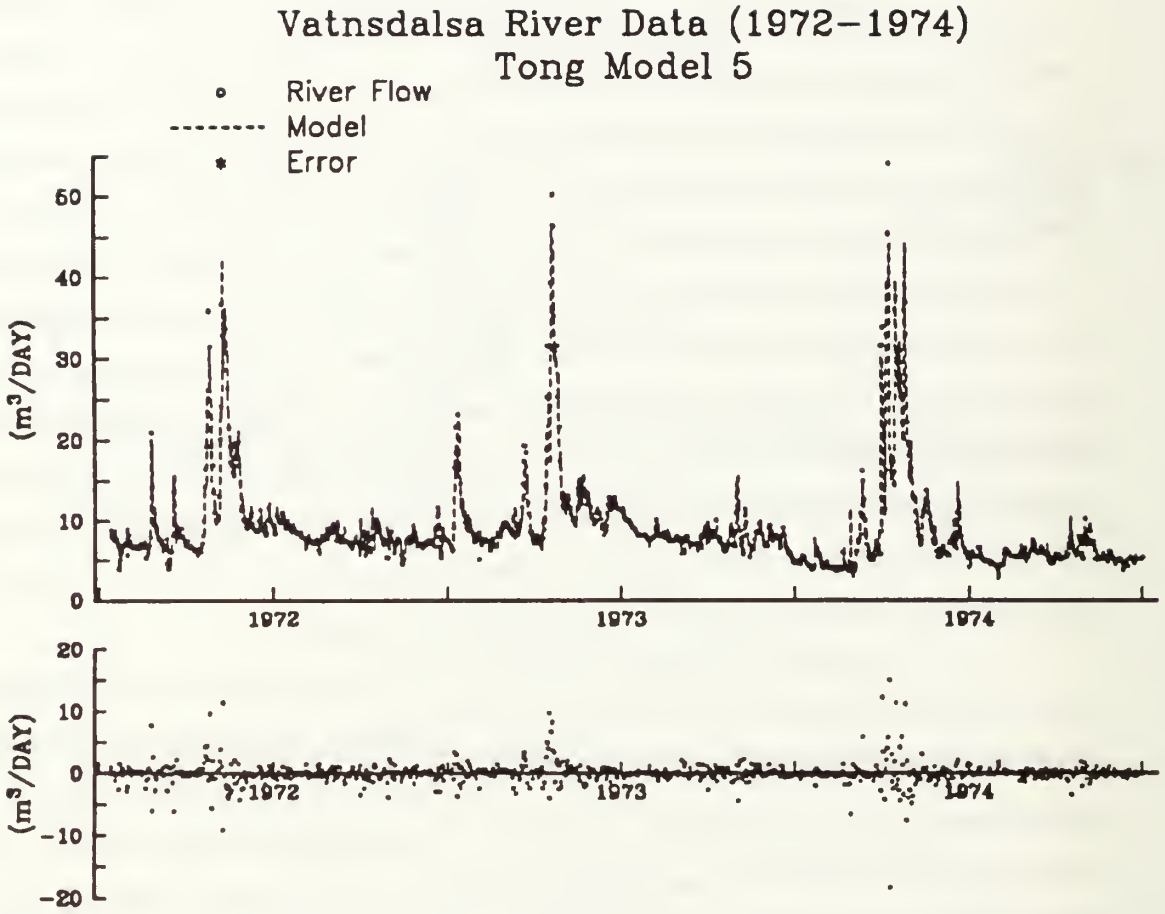

Figure 25. Vatnsdalsa riverflow data for the period 1972-1974 versus the fitted (predicted) values (top) and residuals (bottom) for the final semi-multivariate TAR model, Tong Model 5, from Tong et al. (1985). The semi-multivariate TAR model for the riverflow at time τ, X_τ, is a function of lagged riverflow $X_{\tau-j}$ for $j = 1, \ldots, 10$, and precipitation $Y_{\tau-i}$, and temperature $Z_{\tau-i}$ for $i = 0, \ldots, 10$. The final model contains 73 parameters in 4 disjoint subregions that are described by the 3 temperature thresholds on Z_τ at -2, 2 and 5 $^\circ C$. The standard error of the pooled fitted residuals σ_ϵ is 1.69 m^3/sec. The use of Y_τ and Z_τ in the TAR model essentially prohibits the use of the model for prediction.

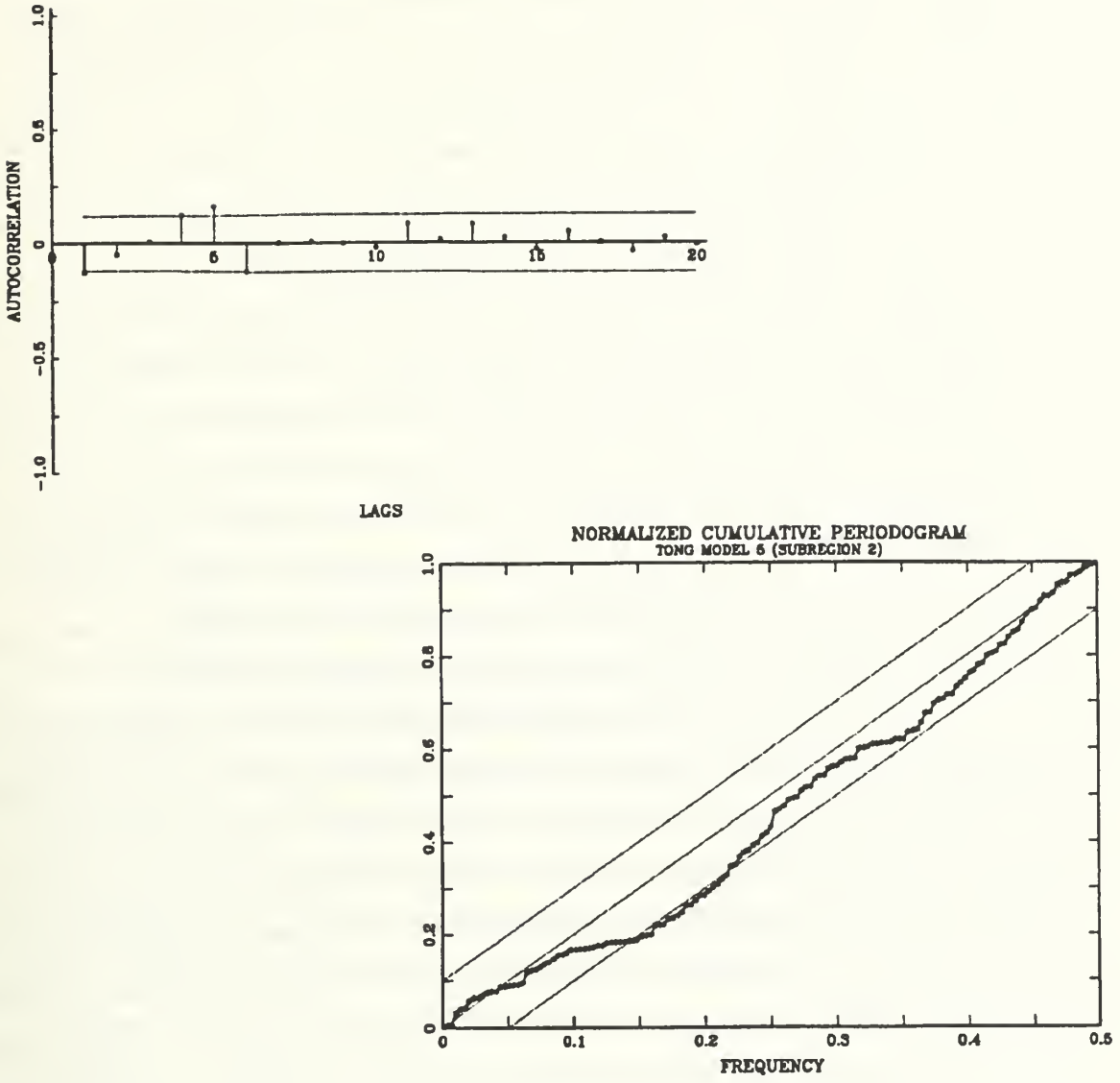

Figure 26. Fitted Residual Plots from Tong Model 5. The autocorrelation function (first 20 lags) and the normalized cumulative periodogram of the fitted residuals from the second subregion, $-2 < Z_T < 2$, of Tong Model 5 from Tong et al. (1985) of the Vatnsdalsa River system for the period 1972-1974. If the fitted residuals of the model are truly Gaussian white noise then the fitted residuals from each subregion should also be independent. Although the fitted values of the model, Figure 25, appear to capture the overall structure of the riverflow data, the approximate 95% individual confidence bounds show that some short term residual autocorrelation in the second subregion still exists. Also, the cumulative normalized spectrum of the fitted residuals falls outside the 90% K-S bounds for Gaussian white noise indicating that we should reject the hypothesis that the fitted residuals are Gaussian white noise. Note that the results of the fitted residual analysis from the other subregions of Tong Model 5 are similar.

TAR models for the Vatnsdalsa riverflow from 1972-1974 using 20 predictor variables; lagged riverflow $X_{\tau-1}$ to $X_{\tau-5}$, lagged precipitation Y_τ to $Y_{\tau-7}$, with and without the natural log transformation $Y^*_{\tau-i} = ln\,(1 + Y_{\tau-i})$, lagged temperature Z_τ to $Z_{\tau-5}$, and a variable for time of year effect. The models were initialized using 9 data values for each of the input time series. The different models obtained occurred because of changes made to the user parameters of the MARS algorithm. These parameters include: MI = 3 and 4, the maximum level of lagged predictor variable interaction; MS = 10, 15 and 20, the minimum separation of a lagged predictor variable's partition points; and M = 30, the number of steps during the forward-step algorithm. The SMASTAR models were identified with the SMASTAR version of MARS 2.0 installed on an IBM3033 Computer using VS Fortran. Each of the 3-year models required from 1 to 2 minutes of CPU time. Also, the maximum lags of each predictor variable time series were chosen because of predictor variable constraints within the SMASTAR version of MARS 2.0. More predictor variables could have been modeled using the adjustments to MARS 3.0 that will be discussed in Chapter IV.

The results of the modeling effort indicate that the SMASTAR methodology appears well suited for analysis of semi-multivariate time series systems. Although it will be discussed in more detail in the next section, 2 and 3 year SMASTAR time series model terms appear to provide an indication of the underlying physical structure of the riverflow system. Throughout the modeling effort it was interesting to note that although the time variable was included as a predictor variable it was *never* selected as a final model term. This in effect, implies that for this riverflow system and data, the lagged predictor variables have captured the relevant time dependent structure of the riverflow. Also, while the (riverflow / precipitation variables) and (riverflow / temperature variables) frequently developed interaction terms in the models, there were few direct interactions between the temperature and precipitation variables. The SMASTAR models developed with the natural log transformation $Y^*_{\tau-i} = ln\,(1 + Y_{\tau-i})$ and a maximum level of interaction of MI = 3, appeared more stable than models developed without the transformation and with MI = 4. As expected, SMASTAR models of the Vatnsdalsa river system are relatively complex when compared to the ASTAR models developed for the yearly Wolf sunspot numbers in Chapter II because interaction terms between cross-correlated predictor variables are permitted.

Figure 27 shows the fitted values and residuals of SMASTAR Model ICE796 of the Vatnsdalsa riverflow for the three years 1972 to 1974. Model ICE796 was selected as a result of the overall fit of the model along with the analysis of its fitted residuals. Model ICE796 was developed using a natural log transformation of the precipitation predictor variable and was permitted to form 1, 2, and 3-way interactions during a maximum of $M = 30$ forward steps of the forward step MARS algorithm. The minimum span between threshold values for a single predictor variable was 15 data values. The model has 37 parameters that include 24 model terms (a constant term and 3 one-way, 8 two-way and 12 three-way interactions) and 13 threshold values (2 for $X_{\tau-1}$, 1 for $X_{\tau-3}$; 2 for $Y_{\tau-1}$, 1 for $Y_{\tau-2}$, 1 for $Y_{\tau-5}$, 2 for $Y_{\tau-6}$; 1 for Z_τ, 1 for $Z_{\tau-1}$, 1 for $Z_{\tau-3}$, and 1 for $Z_{\tau-6}$). It can be seen that Model ICE796 captures the overall structure of the riverflow data. The standard error of the fitted residuals is $\sigma_\epsilon = 1.39 m^3/sec.$

Analysis of the fitted residuals from this model, Figure 28, shows that no short term residual autocorrelation exists in contrast to the short term residual autocorrelation that was present in the TAR models. Also, the residuals could be considered independent if they were normally distributed because the normalized cumulative spectrum of the fitted residuals falls entirely within the 90% K-S bounds from the cumulative spectrum for Guassian white noise. However, the model residuals still appear non-Gaussian with extremely heavy tails that can be expected with this type riverflow data (Figure not shown). Note that SMASTAR Model ICE796 (37 parameters, $\sigma_\epsilon = 1.39 m^3/sec.$), Figure 27, has fewer parameters and smaller fitted residual variance than Tong's TAR Model 5 (73 parameters, $\sigma_\epsilon = 1.69 m^3/sec.$), Figure 25, and also appears to better capture the structure of the periods of high riverflow that occur each spring. Note that the vertical scales of the plots in Figures 25 and 27 are the same.

b. Two Year SMASTAR Models of the Vatnsdalsa River for Prediction

In the previous section we discussed the development of TAR and SMASTAR time series models for three years of the Vatnsdalsa riverflow. However, these TAR and SMASTAR models included same day precipitation and temperature predictor variables, Y_τ and Z_τ, which essentially prohibits the use of the models for riverflow prediction. In this section our objective is prediction; MARS was used to develop SMASTAR models for only 731 days of riverflow and the remaining 355 days were used for prediction. It is unknown if

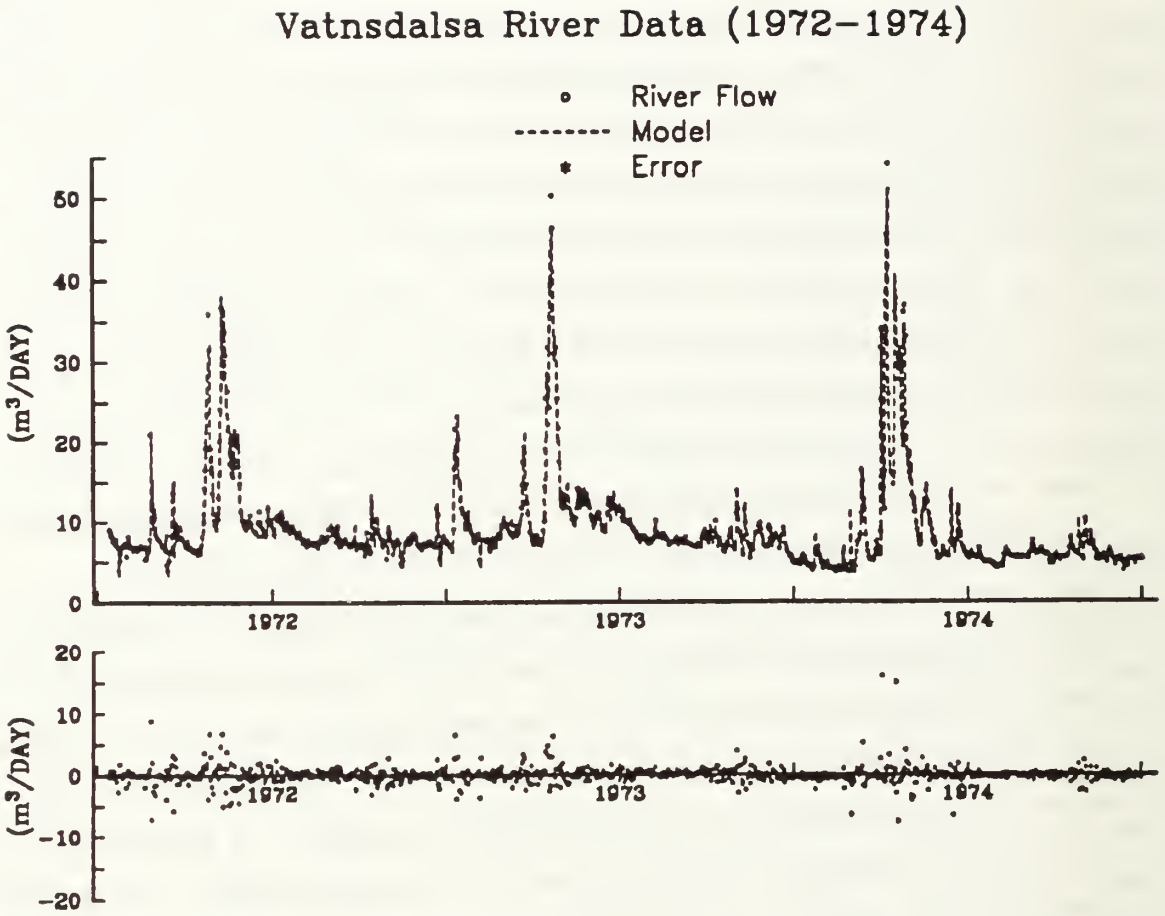

Figure 27. Vatnsdalsa riverflow data versus the fit (top) and the residuals (bottom) for SMASTAR Model ICE796. The period of the modeling effort is 1972 to 1974. The SMASTAR model for the riverflow at time τ, X_τ, is a function of lagged riverflow $X_{\tau-i}$ for $i = 1,\ldots,5$, precipitation $Y_{\tau-j}$ for $j = 0,\ldots,7$, and temperature $Z_{\tau-k}$ for $k = 0,\ldots,5$ and a variable for time of year effect. The final model contains 37 parameters that include 24 model terms (a constant term and 3 one-way, 8 two-way and 12 three-way interactions) and 13 threshold values (2 for $X_{\tau-1}$, 1 for $X_{\tau-3}$; 2 for $Y_{\tau-1}$, 1 for $Y_{\tau-2}$, 1 for $Y_{\tau-5}$, 2 for $Y_{\tau-6}$; 1 for Z_τ, 1 for $Z_{\tau-1}$, 1 for $Z_{\tau-3}$, and 1 for $Z_{\tau-6}$). The standard error of the fitted residuals is $\sigma_\epsilon = 1.39 m^3/sec$.

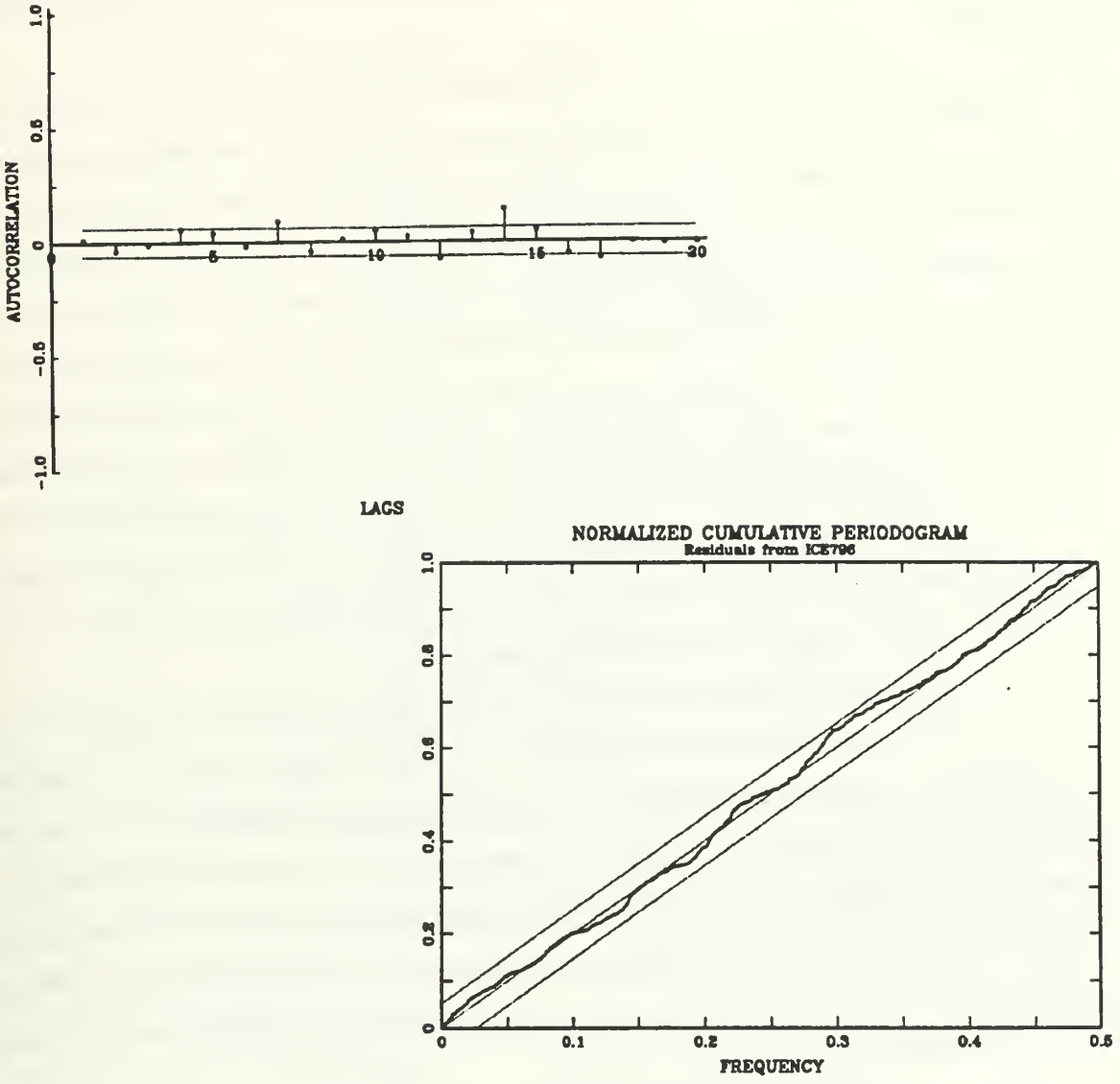

Figure 28. Fitted Residual Plots from SMASTAR Model ICE796. The autocorrelation function (first 20 lags) [top] and the normalized cumulative periodogram [bottom] of the fitted residuals from SMASTAR Model ICE796 of the Vatnsdalsa River system for the period 1972-1974. The autocorrelation plot with approximate 95% individual confidence bounds shows that no apparent autocorrelation exists in the fitted residuals of Model ICE796. Also, the residuals could be considered independent if they were normally distributed because the normalized cumulative spectrum of the fitted residuals falls entirely within the 90% K-S bounds from the cumulative spectrum from Gaussian white noise.

TAR models of the Vatnsdalsa riverflow were developed for purposes of prediction to provide a comparison of the predictive capabilities between SMASTAR and TAR semi-multivariate models.

The SMASTAR models were developed for the 731 days (2 years) of the Vatnsdalsa riverflow during 1972 and 1973 using 20 predictor variables; lagged riverflow $X_{\tau-1}$ to $X_{\tau-5}$, lagged precipitation $Y_{\tau-1}$ to $Y_{\tau-8}$, with and without the natural log transformation $Y^*_{\tau-i} = ln\,(1 + Y_{\tau-i})$, lagged temperature $Z_{\tau-1}$ to $Z_{\tau-6}$, and a variable for time of year effect. The first 9 data values of each time series were used for initialization. Note again that during this modeling effort we excluded Y_τ and Z_τ (same day temperature and precipitation) from the model to permit riverflow prediction for the last 355 days of riverflow during the year 1974. Again different models occurred because of changes made to the user parameters in the MARS algorithm. The parameter selections included: MI = 2, 3 and 4, the maximum level of lagged predictor variable interaction; MS = 10, 15 and 20, the minimum separation of a lagged predictor variable's partition points; and $M = 15$, the number of steps during the forward-step algorithm.

As with the 3-year modeling effort, the 2-year SMASTAR models appear well suited for analysis of semi-multivariate time series systems. Again, the (riverflow / precipitation variables) and (riverflow / temperature variables) frequently developed interaction terms in the models, there were few direct interactions between the temperature and precipitation variables. Also, the SMASTAR models developed with the natural log transformation $Y^*_{\tau-i} = ln(1+Y_{\tau-i})$ and a maximum level of interaction of MI = 3, appeared more stable (less likely to have abnormal changes in riverflow) than models developed without the transformation and with MI = 2 and 4.

Equation (54) details SMASTAR Model ICE486 for the Vatnsdalsa riverflow for the years 1972 and 1973. SMASTAR Model ICE486 was selected from among the other models due to the overall model fit and the analysis of its fitted residuals. The presentation of Model ICE486 in equation (54) is intended to take advantage of the tree-like structure that naturally develops as a result of its truncated spline functions and of the stepwise selection methodology within MARS. Model ICE486 for the Vatnsdalsa riverflow was developed using the natural log transformed precipitation and was permitted to form 1, 2, and 3-way interactions during a maximum of $M = 15$ forward steps of the forward step MARS

algorithm. The minimum span between threshold values for a single predictor variable was $MS = 20$ data values. Model ICE486 is

$$\hat{X}_\tau = \begin{cases} 3.12 & + 2.34(Y^*_{\tau-1} - 2.49)_+ \\ & \\ & + 1.20(X_{\tau-1} - 3.98)_+ \\ & \begin{cases} - .200(X_{\tau-1} - 3.98)_+(.833 - Y^*_{\tau-2})_+ \\ + .038(X_{\tau-1} - 3.98)_+(.833 - Y^*_{\tau-2})_+(14.5 - X_{\tau-4})_+ \\ - .116(X_{\tau-1} - 3.98)_+(Y^*_{\tau-2} - .833)_+ \end{cases} \\ & \\ & \begin{cases} + .174(X_{\tau-1} - 3.98)_+(7.92 - X_{\tau-2})_+ \\ \begin{cases} - .014(X_{\tau-1} - 3.98)_+(X_{\tau-2} - 7.92)_+(3.2 - Z_{\tau-1})_+ \\ - .021(X_{\tau-1} - 3.98)_+(X_{\tau-2} - 7.92)_+(Z_{\tau-1} - 3.2)_+ \end{cases} \\ \begin{cases} + .008(X_{\tau-1} - 3.98)_+(X_{\tau-2} - 7.92)_+(2.4 - Z_{\tau-1})_+ \\ + .012(X_{\tau-1} - 3.98)_+(X_{\tau-2} - 7.92)_+(Z_{\tau-1} - 2.4)_+ \end{cases} \\ \begin{cases} + .008(X_{\tau-1} - 3.98)_+(X_{\tau-2} - 7.92)_+(3.3 - Z_{\tau-2})_+ \\ - .005(X_{\tau-1} - 3.98)_+(X_{\tau-2} - 7.92)_+(Z_{\tau-2} - 3.3)_+. \end{cases} \end{cases} \end{cases} \tag{54}$$

Model ICE486 has 21 parameters that includes 13 terms (a model constant term and 2 one-way, 3 two-way and 7 three-way interactions) and 8 threshold values (1 each on the lagged riverflow predictor variables, $X_{\tau-1}, X_{\tau-2}, X_{\tau-4}$; lagged transformed precipitation variables, $Y^*_{\tau-1}, Y^*_{\tau-2}$; and the lagged temperature predictor variable, $Z_{\tau-2}$, and 2 on the lagged temperature predictor variable, $Z_{\tau-1}$). The standard error of the fitted residuals for the model is $\sigma_\epsilon = 1.27 m^3/sec.$

Figure 29 shows plots of the fitted values and residuals of Model ICE486 for the Vatnsdalsa riverflow data during 1972 and 1973. Again, note that the precipitation data used in Model ICE486 is the natural log transformed precipitation. Model ICE486 appears to capture the overall structure of the Vatnsdalsa riverflow. Note also, that the minimum riverflow for the modeling period is 3.98 $m^3/sec.$, which is higher than the minimum riverflow that occurs during the period we will be using the model for prediction. The size of the 2-year Model ICE486 (21 parameters) and the standard error of the fitted

residuals $\sigma_\epsilon = 1.27 m^3/sec.$ when compared to the 3-year Model ICE796 (37 parameters) with $\sigma_\epsilon = 1.39 m^3/sec$ and 3-year Tong Model 5 (73 parameters) with $\sigma_\epsilon = 1.69 m^3/sec$, provide some insight into the change in riverflow structure that occurs between the first two years (1972-1973) and the last year (1974). The 3-year models (1972-1974) require many more parameters than the 2-year model (1972-1973) to account for the change in riverflow structure during 1974.

Model ICE486, Figure 29, appears to equally overfit and underfit the peaks and troughs as it captures the general structure of the riverflow data. The fitted residuals are examined using the normal probability plot (Figure 30) and the autocorrelation function and estimated normalized periodogram plots (Figure 31). Analysis of the normal probability plot (Figure 30) shows that the fitted residuals are slightly skewed with extremely heavy tails. Note that the heavy tails could be an indication of different distributions for fitted residuals from different regions of the predictor variable space. Again, unlike Tong Model 5 (25) and the other TAR models discussed in the previous section, the autocorrelation function for the fitted residuals reveals no evidence of short term autocorrelation. Also, we could consider the residuals independent if they were normally distributed because the normalized cumulative spectrum of the fitted residuals falls entirely within the 90% K-S bounds from the cumulative spectrum for Gaussian white noise. However, as with these other models, the fitted residuals still display a pattern of high residual values during periods of high riverflow (Figure 29). This is evidence that we have still not captured all the relevant predictor variables for the periods of high level riverflow.

c. *Interpretation of the Two Year SMASTAR Model ICE486*

The tree-like structure of Model ICE486 (54) provides some insight into the complex interactions of the riverflow system. There are three major regions of interest that may be identified by a visual inspection of the equation for the model. They include riverflow when it falls below 3.98 $m^3/sec.$ (top line), along with the model terms that reflect the direct contributions by the lagged transformed precipitation (term 2 of line 1 and lines 3, 4 and 5) and lagged temperature (lines 7 though 12) variables.

Since all terms in (54) that have the term $(X_{\tau-1} - 3.98)_+$ have value 0 when the riverflow falls below 3.98 $m^3/sec.$, it is immediately apparent that Model ICE486, in

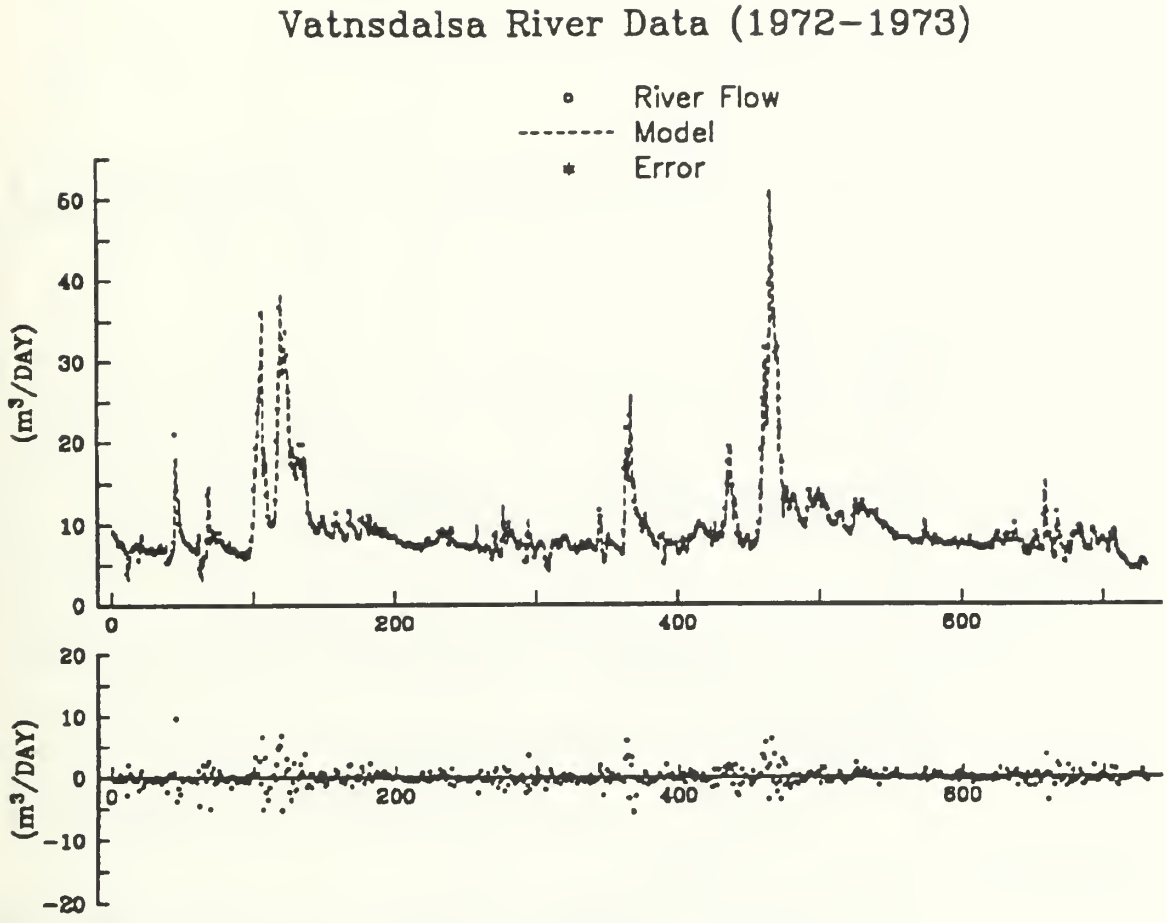

Figure 29. The Vatnsdalsa riverflow data for years 1972 and 1973 versus the fitted values (top) and residuals (bottom) for SMASTAR Model ICE486. The SMASTAR model for the riverflow at time τ, X_τ, was a function of lagged riverflow $X_{\tau-1}$ to $X_{\tau-5}$, lagged precipitation $Y^*_{\tau-1}$ to $Y^*_{\tau-8}$, i.e., the natural log transformation $Y^*_{\tau-i} = ln(1 + Y_{\tau-i})$, lagged temperature $Z_{\tau-1}$ to $Z_{\tau-6}$, and a variable for time of year effect. The final model contains 21 parameters that includes 13 terms with 8 thresholds (1 each on the lagged riverflow predictor variables; $X_{\tau-1}, X_{\tau-2}, X_{\tau-4}, thelaggedprecipitationpredictorvariables; Y_{\tau-1}, Y_{\tau-2}$, and the lagged temperature variable, $Z_{\tau-2}$ and 2 on the lagged temperature variable $Z_{\tau-1}$). The standard error of the fitted residuals σ_ϵ was $1.27 m^3/sec$. The initial nine values of each time series were used to initialize the model.

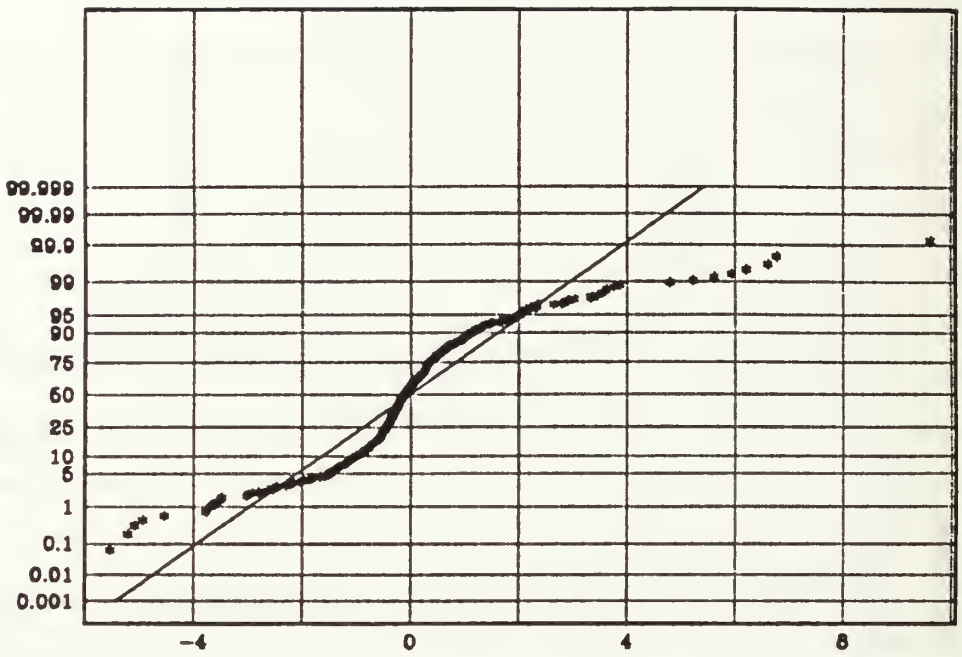

Figure 30. The normal probability plot of the fitted residuals for SMASTAR Model ICE486 of the Vatnsdalsa River system for the period 1972-1974. The horizontal axis shows the range of the fitted residuals from Model ICE486 while the vertical axis shows the corresponding percentiles from the normal distribution. Analysis of this plot shows that the fitted residuals from Model ICE486 are slightly skewed with the extremely heavy tails that we might expect with this type riverflow data. Note that the heavy tails could be an indication of different distributions for the residuals in different regions of the predictor variables.

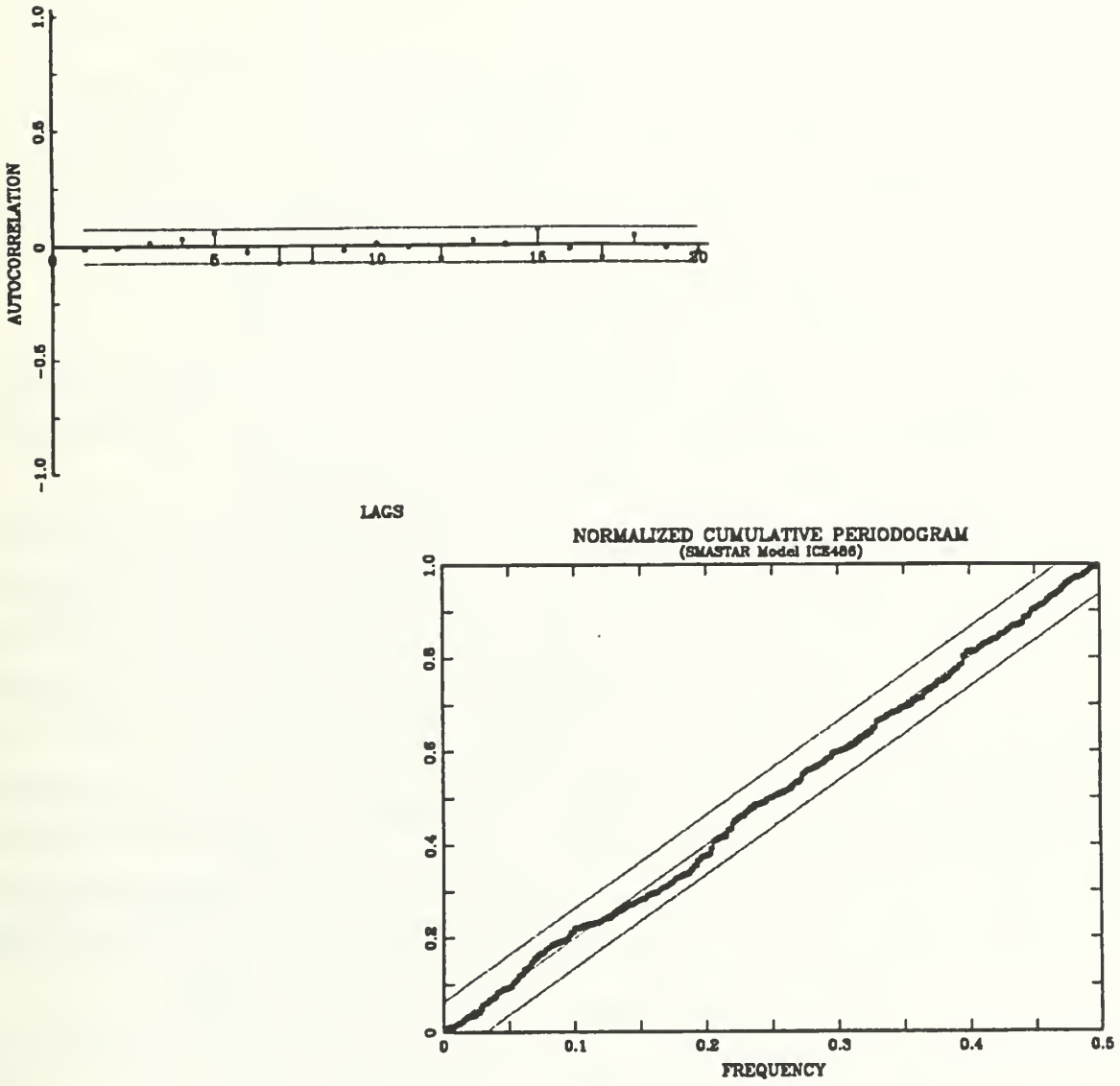

Figure 31. Fitted Residual Plots from SMASTAR Model 486. The autocorrelation function (first 20 lags) [top] and the normalized cumulative periodogram [bottom] of the fitted residuals from SMASTAR Model486 of the Vatnsdalsa River system for the period 1972-1973. The autocorrelation plot with approximate 95% individual confidence bounds shows that no apparent autocorrelation exists in the fitted residuals. Also, we could consider the residuals independent if they were normally distributed because the normalized cumulative spectrum of the fitted residuals falls entirely within the 90% K-S bounds from the cumulative spectrum for Gaussian white noise.

this case, can be reduced to

$$\widehat{X}_\tau = 3.12 \ + \ 2.34(Y^*_{\tau-1} - 2.49)_+. \tag{55}$$

This indicates that when $X_{\tau-1} < 3.98$ and in the absence of lag 1 precipitation $Y^*_{\tau-1}$, the riverflow $\widehat{X}_\tau$ will fall to a steady state level of 3.12 $m^3/sec.$, the model constant. Note that the minimum riverflow during the modeling period was 3.98 $m^3/sec.$ Once the level of riverflow is reduced to 3.12 $m^3/sec.$, then a minimum of $Y^*_{\tau-1} = 2.86$, or $Y_{\tau-1} = 16.42$ millimeters of rainfall must occur to raise the riverflow level above 3.98 $m^3/sec.$, the level at which the other model terms can again 'kick in'. Also, anytime that lag 1 transformed precipitation $Y^*_{\tau-1} > 2.49$ units (or lag 1 precipitation $Y_{\tau-1} > 11.06$ millimeters) there is an immediate contribution to the riverflow as a result of this term.

The next region of interest for Model ICE486 (54) occurs when the lag 1 riverflow $X_{\tau-1}$ is greater than $3.98 m^3/sec.$ and includes the terms of the model that possess lagged transformed precipitation variables (lines 1, 3, 4 and 5 in the equation). These four terms reflect the direct influence of precipitation of the riverflow system. For example, note the positive coefficient for the first precipitation term involving $Y^*_{\tau-1}$ (line 1) and the negative coefficient for the last precipitation term involving $Y^*_{\tau-2}$ (line 5). If significant precipitation occurs ($Y^*_{\tau-1} > 2.49$) there is the immediate (first day) influence of the lag 1 term $Y^*_{\tau-1}$ (line 1) that is moderated the second day by the lag 2 precipitation term $Y^*_{\tau-2}$ (line 5), if the lag 1 riverflow ($X_{\tau-1}$) is greater than 3.98 $m^3/sec.$, i.e., the term on line 5 reflects the decrease in river runoff levels 2 days after a significant rainfall.

The last region of interest includes the last 6 terms of the model. These terms reflect the direct influence of temperature on the riverflow system. The terms include 2 pairs of the lag 1 temperature variable terms $Z_{\tau-1}$ (lines 7 through 10), and 1 pair of lag 2 temperature variable terms $Z_{\tau-2}$ (lines 11 and 12). The threshold values of 2.4, 3.2 and 3.3 ^oC provide the necessary switching mechanisms to correctly modify the changing behavior of the riverflow system as it is affected by temperature. We can use coefficients of these model terms and temperature extremes to characterize the behavior of the model as it is affected by temperature. For example the coefficients for the model terms that are active (making a nonzero contribution) during very low successive days of temperature

$(-.014, .008, .008)$ (lines 7, 9 and 11) and during successive days of rapidly increasing temperatures $(-.021, .012, .008)$ (lines 8, 10 and 11) in effect cancel each other out. Under these conditions, temperature $Z_{\tau-1}$ and $Z_{\tau-2}$ appears to show little direct influence on riverflow. In contrast, during periods of very high successive temperatures $(-.021, .012, -.005)$ (lines 8, 10 and 12) and rapidly falling temperatures $(-.014, .008, -.005)$ (lines 7, 9 and 12) the temperature terms contribute to the model by forcing riverflow to lower levels. All of these results are rapidly identified and seem reasonable.

d. Predictive Performance of SMASTAR Model ICE486

We now investigate the predictive performance of Model ICE486, developed and discussed above. SMASTAR Model ICE486 (54) and the riverflow, precipitation and temperature data during the year 1974 were used to perform a 1 day forward-step ahead predictions of the Vatnsdalsa riverflow. Prediction of this riverflow for this period is a formidable task due the extreme shift in time and magnitude of riverflow that occurs during the spring along with the decrease in riverflow that occurs later in the year during 1974. For example the minimum riverflow during the modeling period was 3.98 $m^3/sec.$, while the minimum riverflow during the prediction period was 3.67 $m^3/sec.$

The prediction effort used two methods; the first method fixes both the model coefficients and model terms (fixed model) as was done in Chapter II using ASTAR Model 9 of the Wolf sunspot numbers to perform forward-step ahead predictions. The second method fixes the model terms and permits daily updating of the model coefficients (coefficient update) using the latest 731 data values of the riverflow system. For example the 1-step ahead prediction of Model ICE486 (54) at each value of τ during 1974 using coefficient update is obtained by first updating the model coefficients using the data $X_{\tau-i}$, $Y_{\tau-i}$ and $Z_{\tau-i}$ for $i = 1, \ldots, 731$ and then making the 1-step ahead prediction. *Updating the model coefficients is just a simple linear regression step because the threshold values of each model term are fixed.* This second method, coefficient update, was implemented to determine what impact changes in riverflow structure during 1974 has on the fixed prediction model and also because of the nonlinear behavior of the system.

Figures 32-34 contain plots of the actual riverflow versus 1-step ahead predictions and the fitted residuals for the Vatnsdalsa riverflow during the year 1974. In both cases the model predictions react very well to both the extreme spring transition and low

riverflow that occurs later in the year. However, as expected the 1-step predictions using coefficient updating prediction (Figure 32) is an improvement over the 1-step fixed model predictions (Figures 33). The standard error of the fitted residuals are σ_ϵ is 2.11 $m^3/sec.$ and σ_ϵ is 2.36 $m^3/sec.$ respectively. Figure 34 gives the estimated normalized periodogram of the fitted residuals from the 1-step ahead predictions of Model ICE486 using the 'coefficient update' prediction model. The cumulative normalized spectrum of the fitted residuals falls outside the 90% K-S bounds for Gaussian white noise thus indicating that the fitted prediction residuals are not Gaussian white noise.

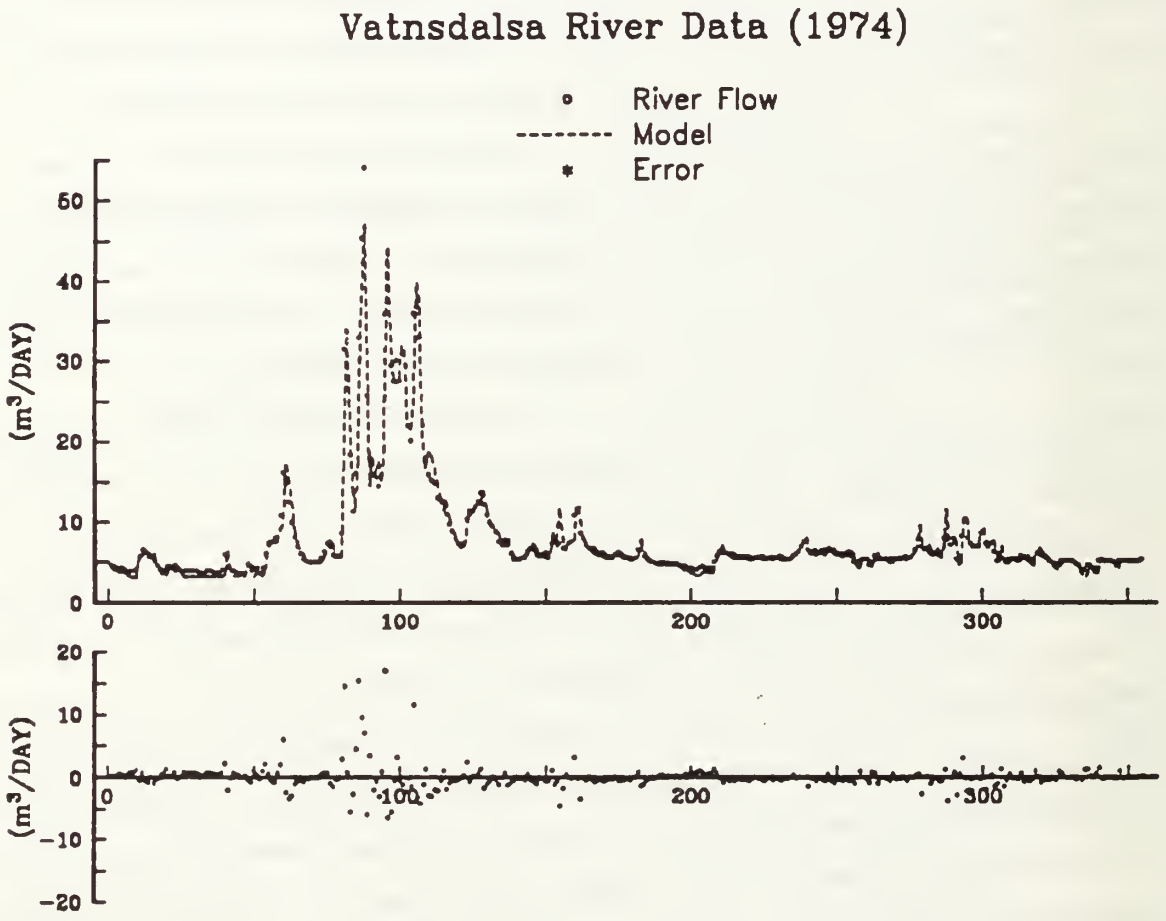

Figure 32. The actual riverflow versus 1-step ahead predictions [top] and errors [bottom] from MODEL ICE486 for the Vatnsdalsa riverflow data (1974) with coefficient updating (coefficient update). The standard error of the fitted residuals σ_ϵ is 2.11 $m^3/sec.$

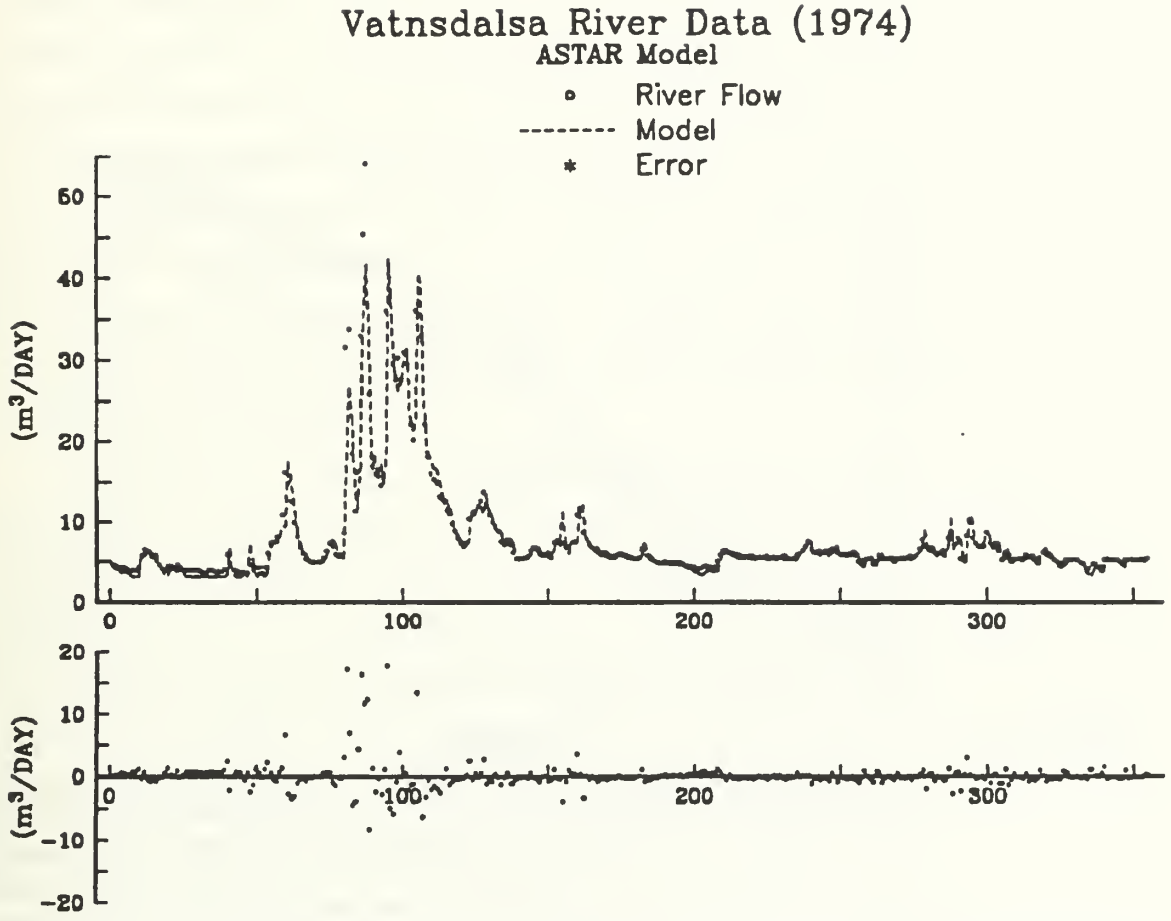

Figure 33. The actual riverflow versus 1-step ahead predictions [top] and errors [bottom] from MODEL ICE486 for the Vatnsdalsa riverflow data (1974) without coefficient updating (fixed model). The standard error of the fitted residuals σ_ϵ is 2.36 m^3/sec.

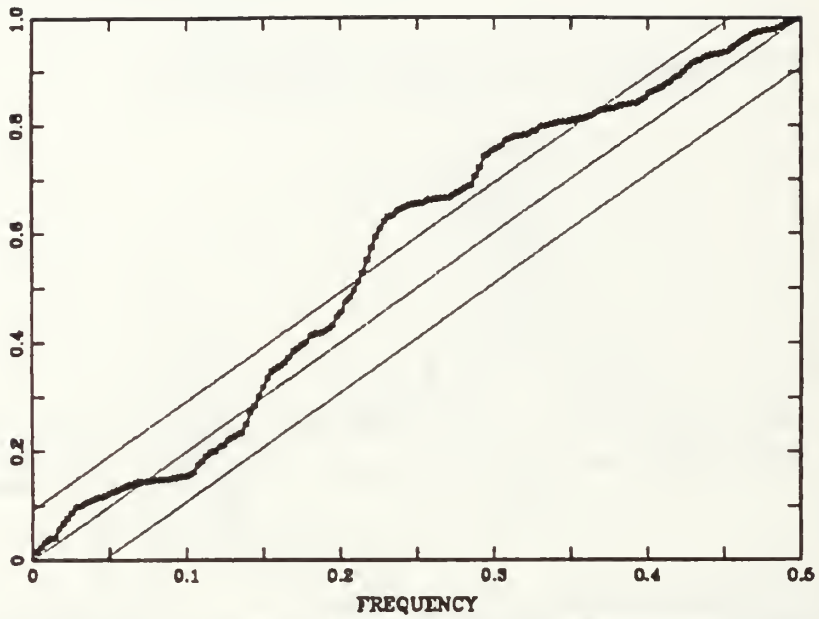

Figure 34. The estimated normalized periodogram of the fitted residuals of SMASTAR Model ICE486 from the Vatnsdalsa riverflow data for 1974 using the 'coefficient update' prediction model. The cumulative normalized spectrum of the fitted residuals falls outside the 90% K-S bounds for Gaussian white noise indicating that we should reject the hypothesis that the fitted residuals from this prediction effort are Gaussian white noise.

C. SUMMARY

This chapter extended the ASTAR modeling methodology developed in Chapter II to semi-multivariate ASTAR (SMASTAR) modeling methodology, which appears well suited for taking into account the complex nonlinear interactions among multivariate, cross-correlated, lagged predictor variables of a time series system. Using the Vatnsdalsa riverflow system as an example, Tong et al. (1985) showed that normal autoregressive models were incapable of capturing the complexities of the cross-correlated predictor variables of this type time series system. Also, the methodology for and structure of semi-multivariate TAR appears incapable of capturing these complexities with a parsimonious model. However, the MARS methodology in the form of a SMASTAR model appears to better consider the complex relationships between the cross-correlated predictor variables and seems capable of providing semi-multivariate nonlinear time series models for prediction. Moreover, the MARS methodology, although computer intensive, provides a systematic approach to modeling time series systems.

It is important to note that the lagged riverflow, precipitation and temperature may only provide rudimentary insights into riverflow modeling and prediction and may not be sufficient for developing a model of this semi-multivariate time series system. Other predictor variables such as wind conditions in the case of the Vatnsdalsa riverflow system may provide important information for modeling of the riverflow system. As with any regression or time series modeling effort, one can never be sure that one has all the relevant predictor variables. However, this additional complexity can be handled in MARS 3.0 with the modifications that will be discussed in Chapter IV.

Other data sets, such as the Canadian Lynx data, and the Sea Surface Temperature data that will be discussed in chapter IV, and many other riverflow data sets exhibit 'periodic' behavior and it would be of interest to model them with the SMASTAR procedure. Of special interest are those data sets with a fixed cycle oscillation that dominates the data. The length of the Vatnsdalsa riverflow data modeled in this chapter may not have been a long enough to satisfactorily establish the fixed yearly oscillation that appears to exist in the SMASTAR models.

IV. MODELING OF TIME SERIES SYSTEMS USING MARS 3.0

The univariate and semi-multivariate ASTAR models developed in Chapters II and III are the result of applying the alpha test version of the MARS 2.0 program (released in December 1989) to the Sunspot numbers and Vatnsdalsa riverflow data sets. Friedman released the alpha test version of the MARS 3.0 program in December 1990. The MARS 3.0 program is a collection of subroutines that implement the multivariate adaptive regression spline strategy developed in Chapter II. Changes in the MARS 3.0 program include plotting subroutines that are useful for interpreting a MARS model, and logistic regression subroutines for modeling categorical variables. Note that these subroutines are of interest but have not been fully investigated for application in a time series setting. The subroutines for use in time series analysis were largely unaffected in the update from the MARS 2.0 program to the MARS 3.0 program.

Our use of the MARS 2.0 program for univariate and multivariate time series modeling and analysis was largely time series specific. For example, our time series modification of the MARS 2.0 program permitted only 20 lagged predictor variables and there was always a residual question as to whether the model would, in some sense, converge if the modeling effort was 'opened up', i.e., if more lagged predictor variables were permitted. Thus, given the results of the ASTAR and SMASTAR time series models developed in Chapters II and III, it was of interest to develop the capabilities of the MARS 3.0 program so that it could be used for the *general modeling and analysis* of any time series system. In particular, the current MARS 3.0 time series program include; simplified input for the program parameters and different input time series, automatic development of the regression matrix for up to three input time series for any combination of lagged predictor variables, automatic computation of memory requirements necessary for the array space calculations used during execution of the MARS 3.0 program, and model output that facilitates analysis of the ASTAR or SMASTAR time series model. In addition, a major change is the inclusion of model selection criteria other than GCV^* (discussed in Chapter V), the original model selection criterion in the MARS 3.0 program. Note that the Fortran Programs presented in

the appendices are for use with NDP Fortran 2.1.4 under DOS with Microway NDP Fortran using the Microway Weitek Coprocessor. This Fortran uses the PharLap DOS Extender to enable Fortran to use all available RAM. However, almost identical programs are available for IBM mainframe computers running VS Fortran.

This chapter is divided into two sections. Section A of this chapter discusses the Fortran programs developed for time series modeling and analysis using the MARS 3.0 program. Section B of this chapter briefly reports on the modeling and analysis of the Granite Canyon sea-surface temperatures using the MARS 3.0 program and the Fortran programs discussed in Section A of this chapter. The sea-surface temperatures are a very long and complex data set with interesting phenomena on many time scales. Thus it is interesting to see how the MARS methodology handles this time series.

A. NEW FORTRAN SUBROUTINES FOR MODELING TIME SERIES SYSTEMS USING MARS 3.0

As presently constituted, the MARS 3.0 program is not simple to use for time series modeling and analysis. The MARS 3.0 program requires that various program parameters be set, does not manage memory requirements for different modeling projects and requires a complete regression design matrix as program input. To overcome these difficulties Fortran programs were developed for time series modeling and analysis using the MARS 3.0 program. Appendices A thru C are Fortran programs to prepare and execute the MARS 3.0 program for the nonlinear modeling and analysis of time series systems. A BATCH program (appendix A) provides useful user information and sequentially executes the MARSBLD (appendix B) and MARSDRV (appendix C) Fortran programs. The BATCH program first calls MARSBLD (appendix B), which asks for the names of up to 3 input time series files and then prepares the regression design matrix and program parameters for input into the MARS 3.0 program. Each time series is located in a separate file with leading lines that contain the model parameters (including the lagged predictor variables) necessary for running the MARS 3.0 program. Next, the BATCH program calls MARSDRV (appendix C), which first computes the memory requirements needed in the MARS 3.0 program and then, if the memory allocation is sufficient, initiates the MARS algorithm described in Chapter II. The

only value that may need adjustment is the parameter for MARS 3.0's memory allocation that is located on the 2nd line of MARSDRV.

The first 3 records of *each* input time series contain the model parameters necessary for initiating the MARS 3.0 program. Note that parameters common to the entire program are marked with an asterisk (*) and are actually taken (read) from the first input time series. The model parameters are;

1. N — The total length of the time series system to be investigated including the initialization values, i.e., only $N - d^*$ values will be modeled where d^* is the maximum lagged predictor variable across all input time series. Each input time series must be of at least length N.
2. P — The total number of predictor variables from the input time series. For example, a time series that is modeled with lags 1, 2 and 10 uses three predictor variables.
3. MI^* — The maximum level (upper bound) of interactions permitted in the generated ASTAR or SMASTAR model. In general, this parameter should be set to $MI \leq 3$. Models permitted to form higher level interactions are difficult to analyze and have a tendency to become unstable.
4. NK^* — The maximum number of steps in the forward-step MARS algorithm. The forward-step algorithm is followed by a backward-step algorithm that trims excess terms from the model.
5. MS^* — The minimum span (in the form of the number of data points) between adjacent thresholds on a lagged predictor variable. This model parameter can be thought of as a smoothing parameter similar to the bandwidth in kernel smoothing. A large value of MS permits fewer threshold values on a given predictor variable.
6. DF^* — The degrees of freedom charged for the selection of a predictor variable, threshold value and coefficient for inclusion in a MARS model. In general values of $2 \leq DF \leq 4$ are recommended with a value of $DF = 3$ used most frequently.
7. MSC^* — The model selection criterion for use within MARS 3.0. The alternatives (discussed in Chapter V) include Friedman's GCV^*, Akaike's AIC, Schwarz and Rissanen's SC, and Amemiya's PC.
8. LX — Flag for each lagged predictor variable of each time series.
   (a) 0 – Directs the predictor variable be excluded from the model.
   (b) 1 – The predictor variable has no restriction. The predictor variable can enter the model with or without a threshold value and also can enter the model as an interaction with other predictor variables.
   (c) 2 – An additive predictor variable. The predictor variable can enter the model with or without a threshold value. However, the predictor variable is not permitted to enter the model as an interaction with other predictor variables.

(d) 3 – A linear predictor variable. The predictor variable can enter the model only as a linear variable, i.e., without an internal threshold. It is not permitted to enter the model as an interaction with other predictor variables.

(e) -1 – A categorical predictor variable with no restriction. The categorical predictor variable can enter the model as an interaction with other predictor variables.

(f) -2 – An additive categorical predictor variable. The predictor variable can enter the model but is not permitted to interact with other predictor variables.

9. *LAGS* — The actual identification of the lagged predictor variables for each time series. For example, *LAGS* = 1, 2 and 5 means the 1st, 2nd and 5th lagged predictor variables. The lagged values must be ordered from smallest to largest. The maximum value of *LAGS* across all input time series will dictate the number of values used for model initialization, d^*.

The sample output (appendix D) of an ASTAR or SMASTAR model that results from the execution of the MARS 3.0 program includes; a summary of the model parameters (discussed above), the forward and backward steps of the MARS algorithm, the final MARS model matrix, the relative benefit of each lagged predictor variable included in the final model and the final model output in a form that permits some model analysis.

B. GRANITE CANYON SEA-SURFACE TEMPERATURES

The Granite Canyon time series is a large data set of the daily raw sea-surface temperatures taken at Granite Canyon, a point just north of Big Sur along the coast of California. Using MARS 3.0 and the Fortran programs discussed in Section A, three ASTAR time series models of this data were developed for test purposes and to compare with results of a previous modeling effort by Breaker and Lewis (1985). The next two parts of this section are a brief background discussion of the Granite Canyon sea-surface temperatures and the modeling effort taken from Breaker and Lewis (1985). The last part of this section is a discussion of three ASTAR time series models of the Granite Canyon sea-surface temperatures developed using the MARS 3.0 program and the Fortran programs discussed in Section A.

1. Sea-Surface Temperatures

Sea-Surface temperatures (SSTs) and their changes in time and space (ocean depth, longitude and latitude) contribute to our understanding of complex ecological issues such as the dispersal of pollutants and fisheries biology. Investigations along the U.S. Pacific Coast indicate that coastal SSTs can be useful indicators of ocean temperature

variability, representative of phenomena occurring over wide regions, related to other ocean and atmospheric variables and have consistent internal structure. Two major factors that contribute to the seasonal variability of SSTs along the California coast include coastal upwelling and the coastal countercurrent. Many investigations of the SSTs appear to focus on temperature anomalies that can persist for several months and influence wide areas of the coast.

Along the California coast, SSTs are collected at approximately 25 locations. The coastal observations often extend over many years and thus provide a unique opportunity to examine coastal variability over relatively long periods. For locations where the measuring site has a good exposure to the adjacent continental shelf and slope, measurements of SSTs may be particularly revealing with respect to some of the physical processes that occur regionally as well as locally. Additional SST background material and references is available in Breaker and Lewis (1988).

2. Spectral Decomposition of the Granite Canyon Sea-Surface Temperatures

The purpose of Breaker and Lewis (1985) was to model the behavior of the 12 year Granite Canyon data set (Figure 35); to use the model and other statistical techniques to project or predict the data to future time and to provide a descriptive interpretation of the Granite Canyon data from the oceanographic viewpoint. The model considered was

$$Y_\tau = M_\tau + S_\tau + \epsilon_\tau \tag{56}$$

where M_τ is a linear trend, S_τ consists of seasonal and cyclic changes, and ϵ_τ is a mean zero, constant variance, stationary random sequence that describes irregular fluctuations and is independent of the other model components.

The modeling procedure initially used least squares regression to identify the linear component $M_\tau = 10.9 + .000374\tau$. There is no doubt that there is an evolutionary trend in the data, probably part of a long term cycle. Unless removed, it corrupts the periodogram with large values at very low frequencies.

Next, after detrending the data with the linear component M_τ, the components S_τ and ϵ_τ were identified using a complex iterative method composed of spectral decomposition and autoregressive time series modeling. The resulting model's long term cyclical and

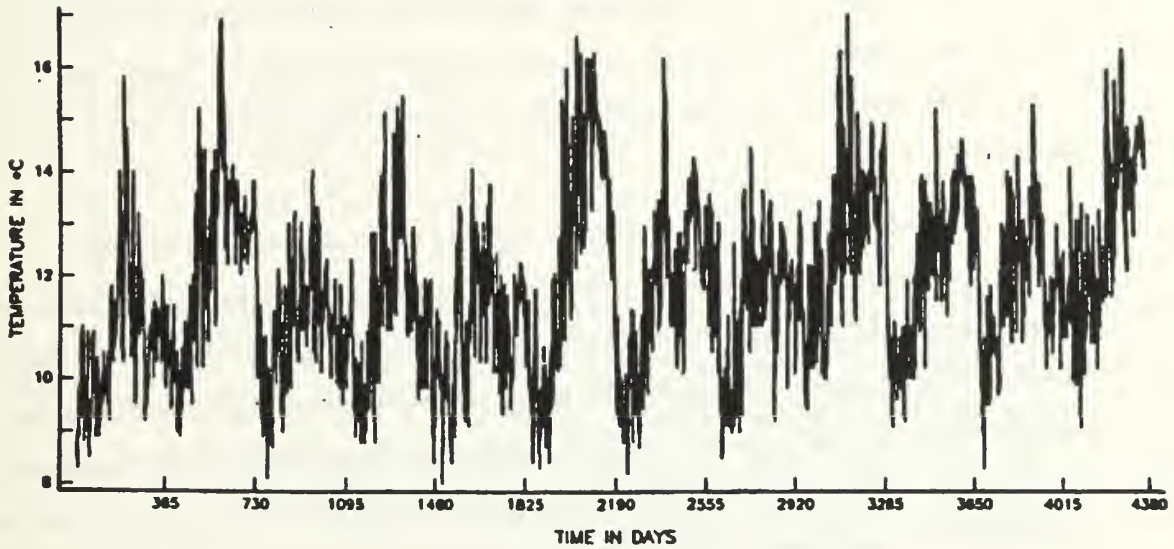

Figure 35. The record of 12 years of daily raw sea-surface temperatures at Granite Canyon from 1 March 1971 to 1 March 1983 taken at approximately 0800 hours each morning. The range of values for the daily sea-surface temperatures range from 8.0 to 17.0 $^{\circ}C$. with a mean value of 11.7 $^{\circ}C$. A least-squares fit of a linear trend to the data indicates that the 'average' temperature rose from about 10.9 to 12.5 $^{\circ}C$ during the 4380 days of the data set. This change in overall temperature level is evident to fishery industry and others as a gross change in the animals and flora seen in the area during this time. Note that the El Nino phenomenon is clearly evident in the record, particularly in 1979 (about day 3200).

seasonal changes component, S_τ, used 13 terms that correspond to 46.6, 182.5, 243.3, 365, 398, 486.6, 547.5, 625.7, 730, 876, 1095, 1460, and 2190 days. There is no physical basis for some of these fixed cycle components; they represent an unsatisfactory edifice for model building. However, there is a definite long term (3 to 5 year) effect of the El Nino, which one would clearly like to model and predict. The error term, ϵ_τ, was modeled as an AR(2) process with a standard deviation of .537 ^oC, although a hump in the correlogram at a lag of about 14 days was unaccounted for. The subsequent analysis of the fitted values and fitted residuals of the model indicated that the model was reasonably adequate and accurate. One and two step predictions (Breaker and Lewis, 1985) of the 30 days immediately following the modeling period (1-30 March 1983) resulted in predictive MSE's of .40 ^oC and .57 ^oC respectively.

3. ASTAR Models of the Granite Canyon Sea-Surface Temperatures

Three ASTAR models of the Granite Canyon sea-surface temperatures were developed using MARS 3.0 and the Fortran programs discussed in Section A. The first model (Granite1) used lags 1 to 49 and lag 365 of the sea-surface temperature series as the model predictor variables; the second model (Granite2) used lags 1 to 50 of the sea-surface temperature series and a discrete valued cosine and sine curve with a period of 1 year as the model predictor variables (the cosine and sine predictor variables were restricted as linear terms i.e., these two predictor variables were not permitted to interact with other predictor variables and were not permitted to form threshold terms); the third model (Granite3) used lags 1 to 50 of the sea-surface temperature series as the model predictor variables. The model parameters were: $MI = 3$, the maximum level of interaction in the ASTAR model; $MS = 50$, the minimum span between threshold values on a predictor variables; $NK = 60$, the number of forward steps in the MARS algorithm with $N = 4380$ days of sea-surface temperatures.

The three ASTAR time series models are similar. Appendix D is the output of the second ASTAR model (Granite2) with lags 1 to 50 and a discrete valued cosine and sine curve with a period of 1 year as the input predictor variables. The model contains 45 terms (a model constant, 5 one-way, 10 two-way and 29 three-way interactions) and 27 threshold values (one on lags 5, 7, 15, 17, 19, 20, 25, 29, 30, 31, 36, 39, 44, 45, and 47; two on lags 2

and 3; three on lag 35 and five on lag 1). The lag 14 and 26 predictor variables enter the model as linear terms, i.e., without interior threshold values. This is interesting because the effect is clearly seen in the correlogram of the detrended data.

Using the relative loss of model fit due to the removal of each term from the model, the most important terms in the model are the lag 1, 2, 14, 36, cosine, 3 and 35 followed by the other terms of the model. The appearance of the lag 26 predictor variable is interesting; it corresponds to the effect, whose origin is as yet unknown, reported on in Breaker and Lewis (1988). ASTAR Model Granite2 is

$$
\hat{X}_\tau = \begin{cases}
15.78 \quad -0.103\ cos(\tau/365) \\
\quad +1.124(X_{\tau-1} - 15.4)_+ - 1.042(15.4 - X_{\tau-1})_+ \\
\quad -0.075(X_{\tau-2} - 8.00)_+ + 0.051(X_{\tau-20} - 9.50)_+ \\
\\
\quad +0.368(15.4 - X_{\tau-1})_+(X_{\tau-2} - 14.9)_+ - 0.214(X_{\tau-2} - 14.8)_+(X_{\tau-14} - 8.00)_+ \\
\quad -0.018(X_{\tau-2} - 8.00)_+(13.4 - X_{\tau-17})_+ - 0.026(X_{\tau-2} - 8.00)_+(X_{\tau-17} - 13.4)_+ \\
\quad -0.159(X_{\tau-2} - 8.00)_+(9.10 - X_{\tau-19})_+ - 0.014(X_{\tau-2} - 8.00)_+(X_{\tau-19} - 9.10)_+ \\
\quad -0.021(X_{\tau-2} - 8.00)_+(12.4 - X_{\tau-36})_+ - 0.356(X_{\tau-3} - 14.8)_+(X_{\tau-14} - 8.00)_+ \\
\quad +0.015(X_{\tau-14} - 8.00)_+(15.4 - X_{\tau-35})_+ - 0.049(X_{\tau-14} - 8.00)_+(X_{\tau-35} - 15.4)_+ \\
\\
\quad -0.018(X_{\tau-1} - 13.1)_+(14.8 - X_{\tau-2})_+(X_{\tau-14} - 8.00)_+ \\
\quad -0.001(15.4 - X_{\tau-1})_+(14.9 - X_{\tau-2})_+(15.0 - X_{\tau-15})_+ \\
\quad +0.057(15.4 - X_{\tau-1})_+(14.9 - X_{\tau-2})_+(X_{\tau-15} - 15.0)_+ \\
\quad +0.006(14.9 - X_{\tau-1})_+(X_{\tau-2} - 8.00)_+(13.4 - X_{\tau-17})_+ \\
\quad +0.054(X_{\tau-1} - 14.9)_+(X_{\tau-2} - 8.00)_+(13.4 - X_{\tau-17})_+ \\
\quad +0.075(10.9 - X_{\tau-1})_+(X_{\tau-2} - 8.00)_+(X_{\tau-36} - 12.4)_+ \\
\quad -0.004(X_{\tau-1} - 10.9)_+(X_{\tau-2} - 8.00)_+(X_{\tau-36} - 12.4)_+ \\
\quad +0.044(X_{\tau-2} - 8.00)_+(X_{\tau-3} - 14.8)_+(X_{\tau-14} - 8.00)_+ \\
\quad +0.026(X_{\tau-2} - 8.00)_+(13.6 - X_{\tau-3})_+(X_{\tau-35} - 13.3)_+ \\
\quad +0.013(X_{\tau-2} - 8.00)_+(X_{\tau-3} - 13.6)_+(13.0 - X_{\tau-45})_+
\end{cases}
$$

$$
\begin{cases}
+0.015(X_{\tau-2}-8.00)_+(X_{\tau-3}-13.6)_+(X_{\tau-45}-13.0)_+ \\
-0.024(X_{\tau-2}-8.00)_+(12.8-X_{\tau-5})_+(X_{\tau-36}-12.4)_+ \\
-0.006(X_{\tau-2}-8.00)_+(11.8-X_{\tau-7})_+(13.4-X_{\tau-17})_+ \\
-0.008(X_{\tau-2}-8.00)_+(X_{\tau-7}-11.8)_+(13.4-X_{\tau-17})_+ \\
-0.031(X_{\tau-2}-14.8)_+(X_{\tau-14}-8.00)_+(X_{\tau-26}-8.00)_+ \\
+0.051(X_{\tau-2}-14.8)_+(X_{\tau-14}-8.00)_+(X_{\tau-36}-8.00)_+ \\
+0.044(14.8-X_{\tau-2})_+(X_{\tau-14}-8.00)_+(X_{\tau-39}-15.0)_+ \\
+0.013(X_{\tau-2}-8.00)_+(X_{\tau-17}-13.4)_+(13.4-X_{\tau-31})_+ \\
+0.015(X_{\tau-2}-8.00)_+(X_{\tau-17}-13.4)_+(X_{\tau-31}-13.4)_+ \\
-0.042(X_{\tau-2}-8.00)_+(X_{\tau-17}-13.4)_+(X_{\tau-44}-14.9)_+ \\
-0.027(X_{\tau-2}-8.00)_+(X_{\tau-19}-9.10)_+(10.0-X_{\tau-35})_+ \\
-0.005(X_{\tau-2}-8.00)_+(X_{\tau-19}-9.10)_+(X_{\tau-35}-10.0)_+ \\
-0.016(X_{\tau-2}-8.00)_+(10.2-X_{\tau-30})_+(12.4-X_{\tau-36})_+ \\
-0.006(X_{\tau-2}-8.00)_+(X_{\tau-30}-10.2)_+(12.4-X_{\tau-36})_+ \\
-0.029(X_{\tau-2}-8.00)_+(X_{\tau-36}-12.4)_+(11.8-X_{\tau-47})_+ \\
-0.019(14.8-X_{\tau-3})_+(X_{\tau-14}-8.00)_+(10.1-X_{\tau-29})_+ \\
-0.003(14.8-X_{\tau-3})_+(X_{\tau-14}-8.00)_+(X_{\tau-29}-10.1)_+ \\
+0.017(X_{\tau-14}-8.00)_+(9.5-X_{\tau-25})_+(15.4-X_{\tau-35})_+
\end{cases}
$$

The results from the three ASTAR models and the spectral decomposition model (56) appear similar. The standard error of the fitted residuals of ASTAR Model Granite2 is $\sigma_\epsilon = .516°C$, versus $.537°C$ for the spectral decomposition model (56). Both models identify the yearly component as an important term along with the importance of lag terms between lag 40 and lag 50. Figures 36 – 39 are plots for the analysis of the fitted residuals of the three ASTAR models. Figure 36 shows the fitted residuals from 1 March 1979 to 28 February 1980 for the three ASTAR models of the Granite Canyon sea-surface temperatures. No pattern appears to exist. Figures 37 – 39 show the histogram, normalized cumulative periodogram and residual probability plots of the fitted residuals from the three ASTAR models for the Granite Canyon sea-surface temperatures. The residuals from the histogram plots are slightly positively skewed. Figure 38 shows that we can consider the residuals independent

if they are normally distributed because the normalized cumulative spectrum of the fitted residuals falls entirely within the 90% K-S bounds from the cumulative spectrum for white noise. However, the residual probability plots in Figure 39 show that the fitted residuals are slightly skewed with heavy tails, thus indicating the nonnormality of the fitted residuals.

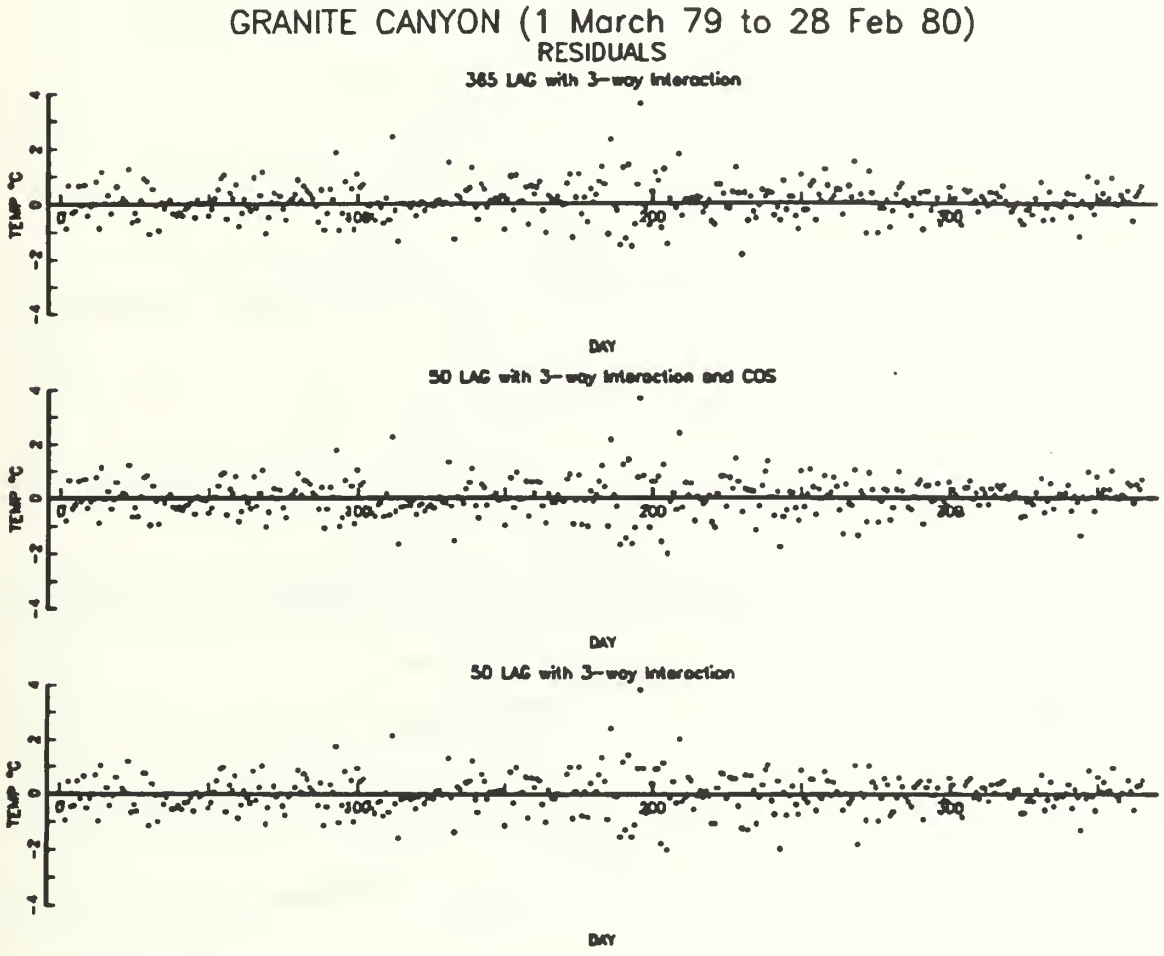

Figure 36. The fitted residuals from 1 March 1979 to 28 February 1980 for three ASTAR time series models of 12 years of daily sea-surface temperatures taken at Granite Canyon. The fitted residuals from each model show no obvious pattern. The fitted residuals from other years are similar.

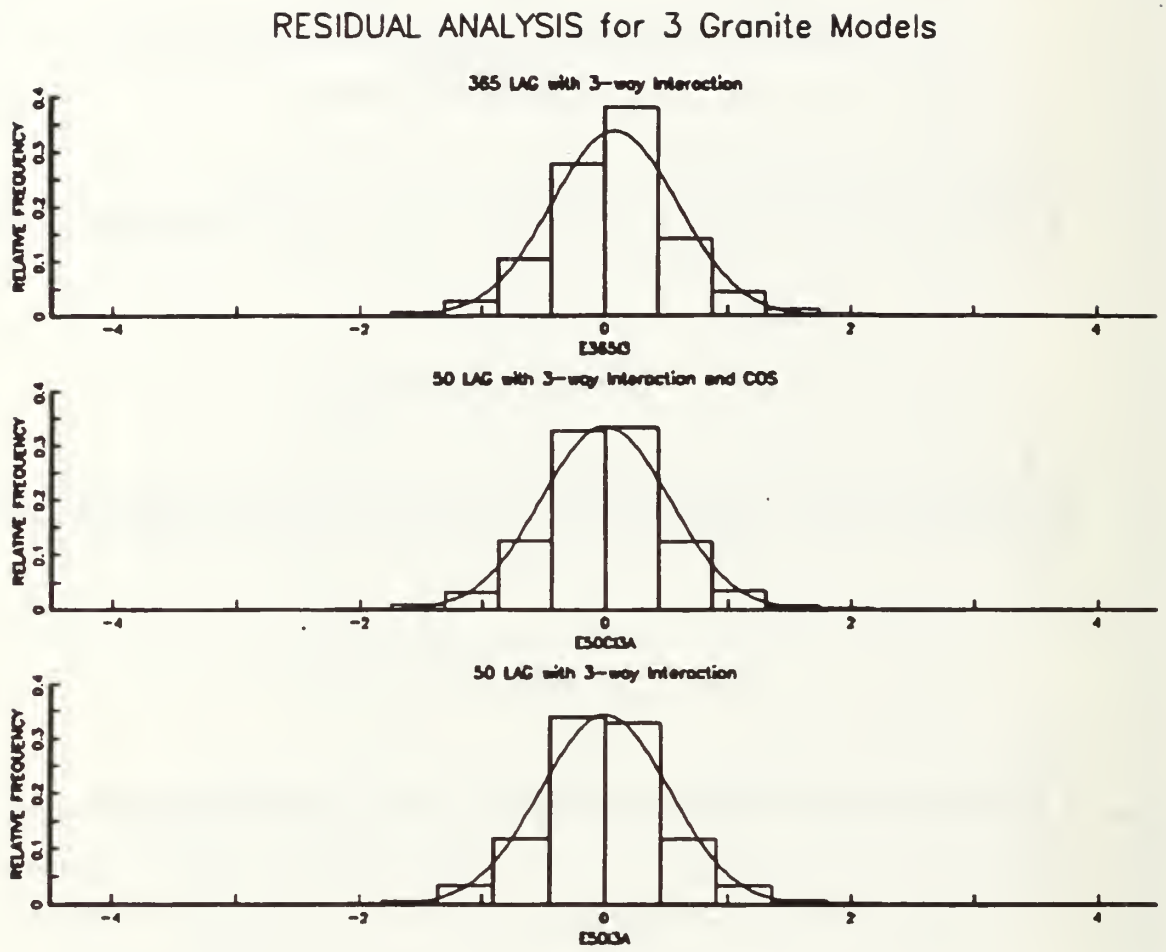

Figure 37. The histogram of the fitted residuals from 12 years of data (March 1971 to 1 March 1980) for three ASTAR time series models of the Granite Canyon sea-surface temperatures. The histograms are overfitted with a normal curve. The fitted residuals from each model appear slightly positively skewed.

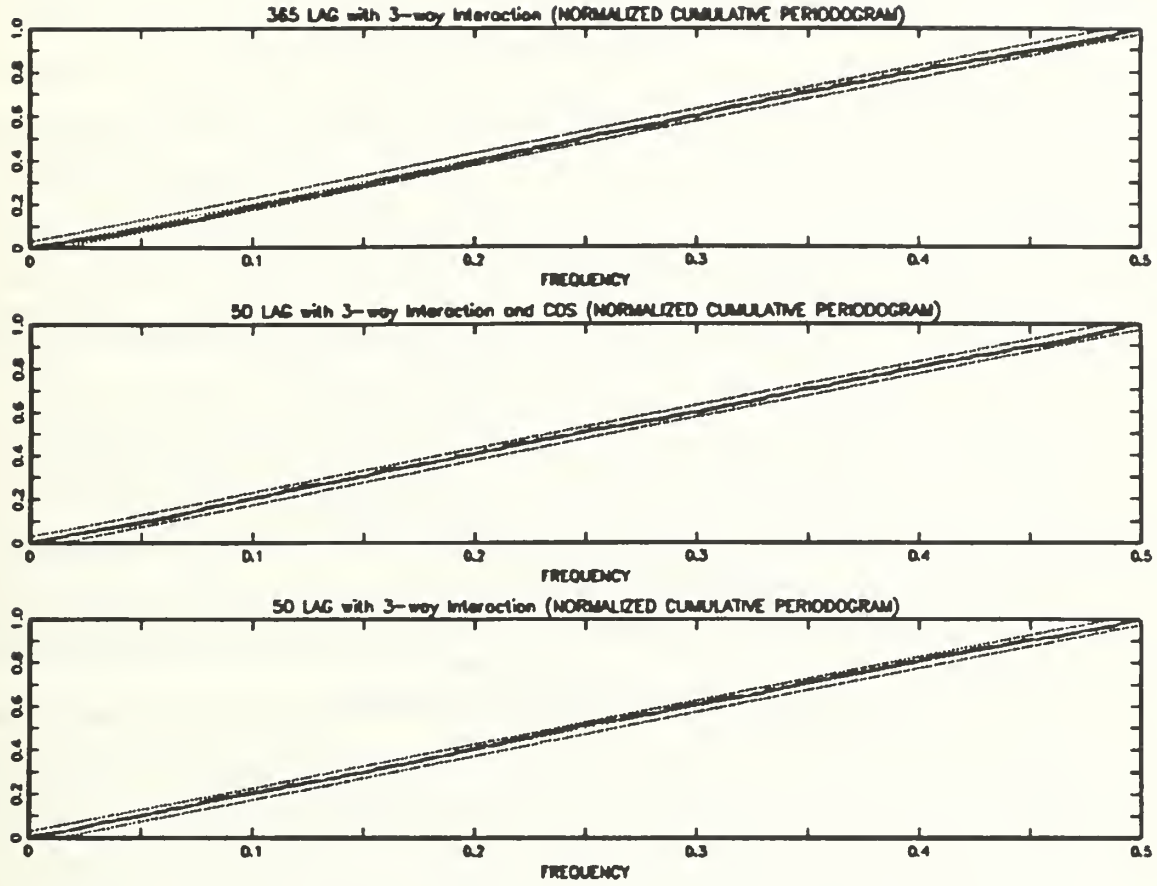

Figure 38. The normalized cumulative periodogram for the fitted residuals from 12 years of data (1 March 1971 to 1 March 1980) for three ASTAR time series models of the Granite Canyon sea-surface temperatures. We can consider the residuals independent if they were normally distributed because the normalized cumulative spectrum of the fitted residuals falls entirely within the 90% K-S bounds from the cumulative spectrum for white noise.

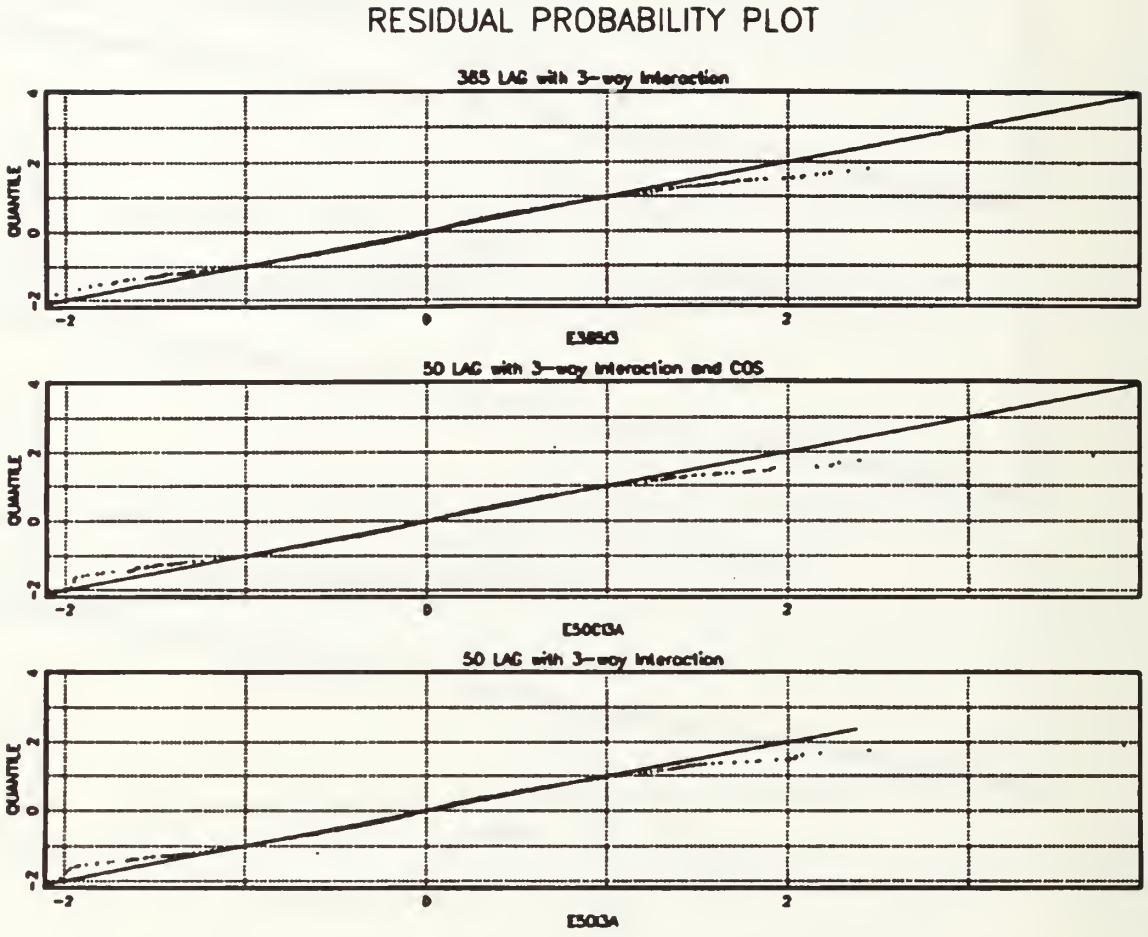

Figure 39. Normal Probability plots of the fitted residuals from 12 years of data (1 March 1971 to 1 March 1980) for three ASTAR time series models of the Granite Canyon sea-surface temperatures. Note that in all three cases the fitted residuals are slightly skewed with heavy tails, thus indicating the nonnormality of the fitted residuals.

C. SUMMARY

The straightforward application of the MARS 3.0 program for the modeling and analysis of time series systems is not simple. This chapter discussed modifications to the MARS 3.0 program and additional Fortran programs that were developed to permit the *general modeling and analysis* of time series systems. Specific changes include; the simplification of program parameter input along with the input of different time series, automatic development of the regression matrix for up to three input time series for any combination of lagged predictor variables, automatic computation of memory requirements necessary for the array space calculations used during execution of the MARS 3.0 program, and model output that facilitates analysis of the ASTAR or SMASTAR time series model. Also, a major change is the inclusion of model selection criteria other than GCV^* (discussed in Chapter V, the original model selection criterion in the MARS 3.0 program.

The MARS methodology in conjunction with these time series modifications represents a new computer intensive but systematic (automatic) modeling approach that isolates the low-dimensional structure among the lagged predictor variables, simplifies the modeling effort and, as shown in Chapters II and III, provides an interpretable representation of a nonlinear time series model that can be used to analyze the relationships between the dependent (output) variable and the independent (explanatory) variables of nonlinear time series systems. The ASTAR time series models for the SSTs generated in this chapter took less than 20 minutes of CPU time on an IBM 3030 mainframe computer using VS Fortran.

Other modeling efforts of sea-surface temperatures have been limited due to the size and complexity of the sea-surface temperature time series system. These limitations appear to be overcome by ASTAR and SMASTAR time series models. What would be of greater interest than the univariate analysis of sea-surface temperatures is an investigation using lagged and cross-correlated sea-surface temperatures, surface winds and time as predictor variables. The application of the MARS algorithm to time series to produce SMASTAR models appears to provide this opportunity.

V. MODEL SELECTION FOR NONLINEAR TIMES SERIES MODELING USING MULTIVARIATE ADAPTIVE SPLINE REGRESSION (MARS)

A. INTRODUCTION

One difficulty that is often faced during the selection of the regression model is the problem of choosing the appropriate predictor (explanatory) variables and model dimension, i.e., which of the given predictor variables to include in the final model, either for the purpose of prediction or for the purpose of description. This chapter examines the problem of model dimension and variable selection when using adaptive regression splines to develop a nonlinear autoregressive model for a univariate or semi-multivariate time series system.

The current MARS algorithm, formulated by Friedman (1991) and implemented in the MARS 3.0 program, uses a form of residual-squared-error as a model selection criterion, in part because of its attractive computational properties. The actual model selection criterion that is used in the forward and backward steps of the MARS algorithm is a modified form of the generalized cross validation criterion (GCV) first proposed by Craven and Wahba (1979). However, one question that immediately develops is whether the modified GCV criterion is the 'best' criterion for model selection when using serially and cross-correlated time series data. Other model selection criteria, such as Akaike's Information Criterion (AIC) (Akaike, 1974), have been suggested for model development in a standard linear autoregressive time series setting.

Section B of this chapter is a brief discussion of five modeling criteria selected for evaluation and comparison within the MARS methodology. The five criteria include GCV^* (Friedman, 1991), Akaike's Information criterion (AIC) (Akaike, 1974) and modified AIC ($AIC2$) (Akaike, 1979), Amemiya's criterion (PC) (Amemiya, 1980) and Schwarz's criterion (SC) (Schwarz, 1978; Rissanen, 1978). Section C of this chapter examines the ability of the different criteria to correctly identify simple linear and nonlinear models and efficiently estimate the model coefficients. However, an approximation to the relationship between

the response variable in terms of the explanatory variables of the time series system may be more important than exact model specification and the exact identification of relationships between the response variable and the different predictor variables. In this regard, Section D of this chapter examines the ability of the different model selection criteria to estimate the fitted values and the limit cycle from ASTAR Model 9 (38) of the Wolf Sunspot numbers. ASTAR Model 9 resulted from the investigation of the ability of MARS to model an actual time series in a more difficult setting. Section E of this chapter is a discussion of the application of the model selection criteria to the Vatnsdalsa riverflow data, where in an 'unrestricted' modeling environment the SC model selection criterion resulted in a better riverflow model than Model ICE486 developed in Chapter III using the GCV^* model selection criterion.

B. MODEL SELECTION CRITERIA

Much of the literature concerned with estimation and inference of a sample time series makes the assumption that we are able to correctly specify the model dimension. However, this situation may be the exception. It is more likely the case that important explanatory variables are omitted or extraneous explanatory variables are permitted in the model. As observed by Akaike (1974), the problem of model selection and fitting in the time series setting is best summarized as a "multiple decision criterion". In this regard, numerous attempts have been made to develop model selection rules and to provide some framework for their use. This section introduces the model selection and fitting problem (Judge et al., 1985) and then briefly discusses the current model selection criterion in MARS, GCV^* (23), and four proposed model selection criteria from linear autoregressive time series modeling for use within MARS; Amemiya's Prediction Criterion (PC) (Amemiya, 1980) and three other 'information theory' based criteria suggested for model selection in a time series setting. The three information theory criteria include Akaike's Information Criterion (AIC) (Akaike, 1974), Schwarz Criterion (SC) (Schwarz, 1978; Rissanen, 1978) and Modified Information Criterion ($AIC2$) (Akaike, 1979). Note that the development and application of the AIC, SC, PC and $AIC2$ criteria are based on the investigation of linear autoregressive and moving average (ARMA) processes. Here, our investigation focuses on the application the AIC, SC, PC and $AIC2$ criteria to non-linear time series processes.

1. Model Selection

A critical aspect in determining the form of the non-parametric regression model during each step of the MARS strategy is the model selection criterion that is used to evaluate model fit and determine the 'proper' model dimension. At each forward step in the MARS algorithm, the model selection criterion is used to select the candidate term that most improves the overall 'goodness-of-fit' for addition to the model. As discussed in Friedman (1991), it follows that at the end of the forward-step procedure there may be model terms that no longer sufficiently contribute to the model fit. Thus at each backward step of the MARS algorithm, the model selection criterion is used to choose a candidate term that least degrades the overall 'goodness-of-fit' for deletion from the model (see Friedman (1991)) for a discussion of the stopping rules for the forward and backward steps of the MARS algorithm).

Without loss of generality assume that MARS is in the backward stepwise procedure, i.e. trimming excess terms from the time series model. As in (1) assume there are N samples of Y and $\boldsymbol{X}$, namely $\{Y_\tau, \boldsymbol{X}_\tau\}_{\tau=1}^{N}$. Using Judge et al., (1985) we can discuss the problem of model selection at a given step in the MARS procedure using the parameterized linear statistical model,

$$\mathrm{Y} = X\beta + \mathrm{e} = X_1\beta_1 + X_2\beta_2 + \mathrm{e}, \tag{57}$$

where Y is the N-dimensional response vector for the model, $X = [X_1, X_2]$ is the current (N x k) design matrix with X_1 and X_2 of dimension (N x k_1) and (N x k_2) respectively, and e is an N-dimensional error vector that has mean zero with variance σ_{e}^2. Also, β is a k-dimensional vector of unknown parameters that is likewise partitioned into parameter vectors β_1 and β_2 of dimension k_1 and k_2 respectively. The least squares estimators of β and σ_{e}^2 are $\mathrm{b} = [b_1, b_2]^{\mathrm{T}}$ and $\hat{\sigma}_e^2$ respectively.

If the model (57) is correct, i.e., the proper dimension of the model is in fact k, then the least squares estimators b and $\hat{\sigma}_e^2$ are minimum variance unbiased estimators of β and σ_{e}^2. Now assume that the matrix X_2 contains the model terms proposed for possible elimination during the backwards step of the MARS algorithm. The question of interest is whether or not to trim 'excess' terms from the model, i.e., whether or not to set

$\beta_2 = 0$. Eliminating necessary terms from the model (setting $\beta_2 = 0$) results in the least squares estimators b and $\hat{\sigma}_e^2$ and the estimates of y being biased while failing to eliminate unnecessary terms from the model (setting $\beta_2 \neq 0$), results in the least squares estimator b and the estimates of y having increased variance (roughness) (Rao, 1971). Thus the question of whether or not to set $\beta_2 = 0$ leads an implicit or explicit determination of the tradeoff between the conflicting objectives of bias and variance. One approach for comparing this trade-off and determining whether or not to set $\beta_2 = 0$ is overall Mean Square Error (MSE).

MSE has been used as the basis of development for many model selection criterion such as Mallows' C_p (Mallows, 1973) and the GCV^* (Friedman, 1988) and PC (Amemiya, 1980) model selection criteria investigated in this chapter. The form of these model selection criterion and the others investigated in this chapter may be divided into two distinct parts, one part that considers lack-of-fit between the proposed model and data (most frequently a function of the residual sum of squares) and the other part that considers model complexity (usually a function of the number of independent parameters in the model). Adding additional terms to a regression model permits a decrease in the model's lack-of-fit that incurs a corresponding increase in model complexity. The model that minimizes a given model selection criterion across all investigated models is selected as the 'best' regression model. Note that all of the model selection criterion investigated in this chapter are some form of a modification of the GCV^* criterion, the current model selection criterion in MARS, and are easily (though tediously) incorporated into the MARS program. This modification of the implementation of the MARS 3.0 program was discussed in Chapter IV. Another model selection criterion for which major modifications would be required of MARS and was therefore not considered is Parzen's CAT criterion (1974).

2. Modified Generalized Cross Validation (GCV^*)

The model selection criterion that is currently used for model selection in MARS is a modified form of the generalized cross validation criterion (GCV) first proposed by Craven and Wahba (1979). GCV was developed as an extension of the cross validation (CV) criterion pioneered by Stone (1977). Both Craven and Wahba (1979) and Friedman (1991) provide discussion and references for the development and use of the GCV criterion.

If we let the residual sum of squares between the data and the fitted model be

$$\hat{\sigma}_\epsilon^2 = \sum_{i=1}^{N} [y_i - \hat{f}(\boldsymbol{x}_i)]^2, \tag{58}$$

then the modified generalized cross validation criterion (GCV^*) used in a MARS model with subregions $\{R_j\}_{j=1}^{M}$ is,

$$GCV^*(M) = \left(\frac{\hat{\sigma}_\epsilon^2}{N}\right)\left(\frac{1}{[1 - \frac{C(M)^*}{N}]^2}\right). \tag{59}$$

Again as discussed in Chapter II the difference between GCV^* and GCV is in the computation of $C(M)^*$, a model complexity penalty function that is increasing in M, the number of nonconstant basis functions in the MARS model (Friedman, 1991). $C(M)^*$ is representative of the number of independent model parameters in a MARS model with M subregions, and accounts for the heavy use of the data in determining both the predictor variables and the predictor variable partition points in addition to the usual model coefficients. Typically the residual sum of squares decreases as the model becomes more and more complex, but the second term increases so that at some point a minimum is reached.

Friedman (1991) provides valuable insights into the use of the GCV^* criterion for various types of MARS modeling. However, the setting that Friedman proposes for the use of the GCV^* criterion does not assume serial correlation among the predictor variables. Thus there is a question whether the GCV^* criterion is the "best" criterion within MARS for the development of ASTAR and SMASTAR models using serial correlated and cross-correlated predictor variables.

3. Model Selection using Information Theory

Many of the popular model selection criterion that are used in a linear times series setting are based on information theory. Most are an outgrowth of the development of the AIC criterion, which is based on the Kullback-Leibler Information Criterion (Akaike, 1974). The objective of a model selection criterion that is based on information theory is to select a model that 'best' incorporates the conflicting considerations of precision of the model estimates (again a measure of the remaining lack-of-fit of the model) and model parsimony (usually a measure of model complexity).

a. Akaike's Information Criterion (AIC)

The use of AIC as a model selection criterion is popular because of its simplicity. However, there are some indications that the AIC Criterion, in the context of linear autoregressive time series modeling, overestimates the number of model parameters, thus favoring a decrease in model lack-of-fit with respect to model complexity i.e., the AIC criterion develops an over-parameterized model. The AIC criterion for a MARS model with subregions $\{R_j\}_{j=1}^{M}$ is,

$$AIC(M) = ln\left(\frac{\hat{\sigma}_\epsilon^2}{N}\right) + 2\left(\frac{C(M)^*}{N}\right). \tag{60}$$

Note that

$$ln\,(GCV^*(M)) = AIC(M) + 2\left(ln\left(1 + \tfrac{C(M)^*}{N-C(M)^*}\right) - \left(\tfrac{C(M)^*}{N}\right)\right)$$

Using the first three terms of a Taylor series expansion to approximate

$$1 + \frac{C(M)^*}{N - C(M)^*}$$

gives,

$$\begin{aligned} ln\,(GCV^*(M)) &= AIC(M) + 2\left(ln(1) + \tfrac{C(M)^*}{N-C(M)^*} - .5(\tfrac{C(M)^*}{N-C(M)^*})^2 - \left(\tfrac{C(M)^*}{N}\right)\right) \\ &= AIC(M) + 2\left(\tfrac{(C(M)^*)^2}{N(N-C(M)^*)} - .5(\tfrac{C(M)^*}{N-C(M)^*})^2\right) \\ &= AIC(M) + o\left(\tfrac{1}{N}\right), \end{aligned}$$

so that the AIC and GCV^* criteria are closely related, especially when N, the sample size, is large.

b. Schwarz Criterion (SC)

In response to indications that the AIC criterion over-parameterizes the model, Schwarz (1978) developed a model selection criterion using a Bayesian argument. At the same time, Rissanen developed (see Rissanen, 1987) a model selection criterion using stochastic complexity analysis to evaluate the uncertainty in the data. When applied to linear time series modeling Risssanen's criterion is equivalent to the Schwarz criterion. Note

that Rissanen (1987) makes a strong case for the use of this criterion because of its apparent widespread applicability. In comparison to the AIC criterion (60) the Schwarz-Rissanen (SC) criterion increases the penalty for adding additional terms to the model by a factor of $(1/2)\ ln(N)$. The SC criterion for a MARS model with subregions $\{R_j\}_{j=1}^{M}$ is,

$$SC(M) = ln\left(\frac{\hat{\sigma}_\epsilon^2}{N}\right) + \frac{ln\,(N)\ C(M)^*}{N}. \tag{61}$$

c. Akaike's Bayesian Information Criterion ($AIC2$)

Akaike (1979) also used a Bayesian framework to develop a criterion for selecting a more parsimonious linear time series model than the AIC criterion (60), i.e. a criterion that like the SC criterion (61) increases the importance of model complexity with respect to the model lack-of-fit within the regression model. The $AIC2$ criterion for a MARS model with subregions $\{R_j\}_{j=1}^{M}$ is,

$$AIC2(M) = (N - C(M)^*)\ ln\left(\frac{\hat{\sigma}_\epsilon^2}{N - C(M)^*}\right) + C(M)^*\ ln\left(\frac{\left(\sum_{i=1}^{N} y_i^2\right) - \hat{\sigma}_\epsilon^2}{C(M)^*}\right). \tag{62}$$

4. Amemiya's Prediction Criterion (PC)

To consider the cost associated with selecting an incorrect model, Amemiya (1980) developed a model selection criterion based on minimizing the unconditional mean squared prediction error. This results in a modification to the AIC criterion (60) that corrects for increasing complexity due to adding additional terms to the MARS model. Again, as with the SC (61) and AIC2 (62) criteria, Amemiya's PC criterion imposes a heavier penalty than the AIC criterion (60) for adding additional terms to a model. Amemiya's PC criterion for a MARS model with subregions $\{R_j\}_{j=1}^{M}$ is,

$$PC(M) = \left(\frac{\hat{\sigma}_\epsilon^2}{N - C(M)^*}\right)\left(\frac{N + C(M)^*}{N}\right). \tag{63}$$

Note that

$$\begin{aligned} GCV^*(M) &= PC(M) + \left(\tfrac{C(M)^*}{N - C(M)^*}\right)^2 \\ &= PC(M) + o\left(\tfrac{1}{N}\right), \end{aligned}$$

so that the PC and GCV^* criteria are closely related, especially when N, the sample size, is large.

C. SOME SIMPLE SIMULATIONS TO COMPARE MODEL SELECTION CRITERIA

In Chapter II simulations were used to show the ability of MARS to identify and estimate the coefficients of simple linear and nonlinear time series models. In this section simulations are now used to initially examine the relative ability of the model selection criteria discussed in Section B, to identify and estimate the coefficients of simple linear and nonlinear time series models. Again, the simulation of an AR(1) model with known coefficients examines the relative ability of each model selection criterion to detect and model a simple linear time series within the framework of the MARS methodology. The simulation of a threshold model with 'AR(1) - like' models in each disjoint subregion examines the relative ability of each model selection criterion to detect and model simple nonlinear threshold time series within the framework of the MARS methodology. As in Chapter II the interest in these simulations is two-fold: how often was the true model identified by each model selection criterion and if not, did the model selection criterion overestimate or underestimate the number of model parameters. Secondly, if the true model was identified how well were the model parameters estimated by each model selection criterion.

1. AR(1) Time Series Model Simulations

As in Chapter II the initial simulation experiment uses the AR(1) model,

$$X_\tau = \rho X_{\tau-1} + K + \epsilon_\tau \tag{64}$$

where $\tau = 1, 2, \ldots, N$ indexes the time series, ρ is a constant coefficient, K is the model constant taken to be zero, and ϵ_τ is $N(0, \sigma_\epsilon^2)$. As described in Chapter II the model is usually considered under the stationarity conditions ($| \rho | < 1$), but non-stationary processes such as random walks ($| \rho | = 1$) and explosive processes ($| \rho | > 1$), are also of interest.

Again, two categories of experiments were conducted using the AR(1) time series model (equation 64).

The first experiment required the model selection criterion within MARS to estimate an AR(1) time series model from the simulated data using one lag predictor variable $X_{\tau-1}$, and using $M = 3$, the maximum number of subregions in the forward-step MARS procedure. The alternative models for the first experiment (to the AR(1) time series model) either have no $X_{\tau-1}$ term (a constant model) or have an $X_{\tau-1}$ term with an internal threshold value t greater than $\min\{X_{\tau-1}\}_{\tau=1}^{N-1}$.

The second experiment required the model selection criterion within MARS to estimate an AR(1) time series model from the simulated data when up to four lag predictor variables, $\{X_{\tau-i}\}_{i=1}^{4}$, are allowed and using $M = 8$, the maximum number of subregions allowed in the forward-step MARS procedure. The alternative models for the second experiment include constant models, nonlinear time series models with at least one internal threshold value, and any time series model that includes a term other than $X_{\tau-1}$, i.e., lags $\tau-2$, $\tau-3$, or $\tau-4$.

Simulation experiments were performed for various combinations of ρ and for various values of the smoothing parameter MS, the minimum number of data points between knots on the same predictor variable. Table 6 and Figures 40–45 show the simulation results for $\rho = .5$, .7, and .9 using $\sigma_\epsilon^2 = N(0,1)$, with a smoothing parameter of MS $= .02N$ data points. Table 6 shows the number of simulations correctly identified as AR(1) time series models by each model selection criterion out of the 100 simulated AR(1) models for a given length of the simulated time series N. On the left and right side of the table are the results of the first and second experiments in which MARS attempted to identify the AR(1) time series model (64) from the simulated data using $P = 1$ (left) and $P = 4$ (right) lagged predictor variables and using $M = 3$ (left) and $M = 8$ (right), the maximum number of subregions allowed in the forward-step MARS procedure.

Overall, the SC criterion performs the best at correctly identifying the simulated data as the AR(1) time series model for all values of N, while the number of correctly identified models using GCV^*, PC and $AIC2$ improves for increasing values of N and becomes comparable to the performance of the SC criterion. The number of models correctly identified by AIC is low throughout the simulation experiment. Further investigation indicates that most of the incorrectly identified models developed by AIC included additional model terms, i.e., as discussed in Section B, AIC appears to overestimate the number of model

TABLE 6. AR(1) MODEL SIMULATION: The number of AR(1) simulations correctly identified by each model selection criterion within MARS for increasing values of N. The model parameters are $\rho = .5$ (top), .7 (middle) and .9 (bottom), and $K = 0$, with $\sigma_\epsilon^2 = N(0,1)$, and a minimum span of MS = $.02(N)$ data points between knots. MARS attempted to identify the AR(1) model (64) from the simulated data using $P = 1$ (left) and $P = 4$ (right) lagged predictor variables and using $M = 3$ (left) and $M = 8$ (right), the maximum number of subregions allowed in the forward-step MARS procedure. Each simulation consists of 100 replications. Overall, SC is the best model selection criterion for correctly identifying the AR(1) simulations. Also, the number of correctly identified models by GCV^*, PC and $AIC2$ improves for increasing values of N. The number of correctly identified models by AIC is low throughout the simulation experiment.

	$P = 1$ and $M = 3$				$P = 4$ and $M = 8$			
N	100	250	500	750	100	250	500	750
	$\rho = .5$							
GCV^*	55	85	94	99	38	56	76	89
AIC	61	60	60	56	42	42	49	40
PC	61	85	92	99	38	54	75	87
SC	81	96	97	97	81	96	97	97
$AIC2$	67	75	85	86	45	63	80	84
	$\rho = .7$							
GCV^*	63	81	98	99	36	51	86	91
AIC	62	58	49	56	47	43	39	41
PC	69	82	96	99	34	55	84	90
SC	92	96	99	97	92	96	99	97
$AIC2$	83	84	91	93	77	80	90	92
	$\rho = .9$							
GCV^*	70	82	95	98	32	53	82	94
AIC	55	54	50	50	47	39	44	42
PC	74	81	94	98	29	50	85	91
SC	94	90	97	94	93	90	97	94
$AIC2$	95	87	97	94	94	87	97	94

parameters. For example in the case that AIC was used to identify the AR(1) time series model with $\rho = .5$ (top of Table 6) from the simulated data using $P = 1$ and $M = 3$, only one of the simulations for $N = 100$ was identified as a constant model, the other 44 misidentified models included at least an internal threshold value. This may be preferable to not identifying any structure at all. A model identified using the AIC criterion in this experiment may still closely approximate the output of the true underlying AR(1) time series model. Nevertheless, this is an indication that the AIC criterion over-parameterizes a proposed model.

Figures 40–45 are a series of box plots for the estimated coefficients of the simulation models correctly identified as AR(1) time series models (as addressed in Table 6) by each model selection criterion within MARS for increasing values of N. For each value of N the boxplots represent the estimated model coefficients using, from left to right, GCV^*, AIC, PC, SC and $AIC2$. The estimates for $\hat{\rho}$ are given in the top set of boxplots, and the estimates for $\widehat{K}$ are given in the bottom set of boxplots. The true value of the model coefficients, $\rho = .5$ (Figures 40–41), $\rho = .7$ (Figures 42–43), and $\rho = .9$ (Figures 44–45), along with $K = 0$ are identified by the dashed line across each of the boxplots. At the bottom of each boxplot is the length N of each simulated time series. By comparing the true values of the model coefficients and the boxplots of the estimated values of the model coefficients across increasing values of N, it is observed that the estimated coefficient values for each of the model selection criterion tend to the true value as N increases.

2. Nonlinear Threshold Time Series Model Simulations

To observe the ability of each model selection criterion within MARS to capture nonlinear threshold model characteristics, simulation of the 2-subregion threshold model (Tong, 1983)

$$X_\tau = \begin{cases} \rho_1 X_{\tau-1} + \epsilon_\tau & \text{if } X_{\tau-1} \leq t \\ \rho_2 X_{\tau-1} + \epsilon_\tau & \text{if } X_{\tau-1} > t \end{cases} \tag{65}$$

was considered, where $\tau = 1, 2, \ldots, N$ indexes the time series, ρ_1 and ρ_2 are constant coefficients, $t = 0$ and ϵ_τ is $N(0, \sigma_\epsilon^2)$. As in Chapter II note that the nonlinear threshold time series model (65) has an 'AR(1)-like' model in each subregion, which implies that, with

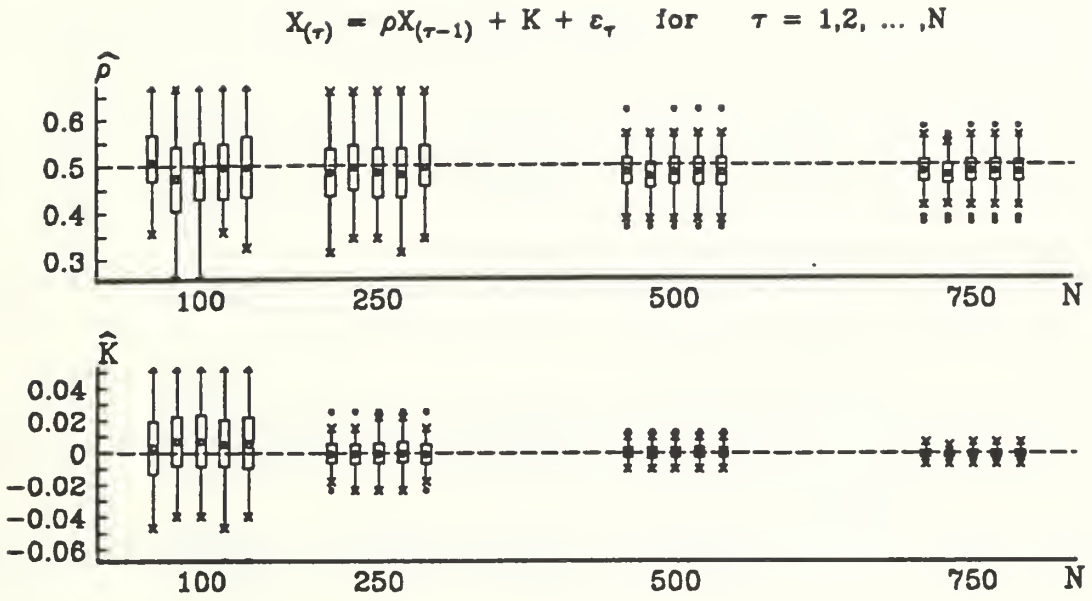

Figure 40. AR(1) MODEL SIMULATION: Boxplots of the estimates from each model selection criterion for $\rho = .5, K = 0$ when the model selection criterion within MARS correctly identified the data as from an AR(1) model (as reflected in Table 6). For increasing values of N, MARS attempted to identify the AR(1) model (64) from the simulated data using $P = 1$ lagged predictor variables and $M = 3$, the maximum number of subregions allowed in the forward-step MARS procedure, with $\sigma_\epsilon^2 = N(0,1)$ and a minimum span of MS $= .02(N)$ data points between knots. Each simulation consists of 100 replications. The model selection criterion represented by the boxplots are, from left to right and for each value of N; GCV^*, AIC, PC, SC and $AIC2$. The true value of the model coefficients, $\rho = .5$ and $K = 0$, are identified by the dashed line across each of the boxplots.

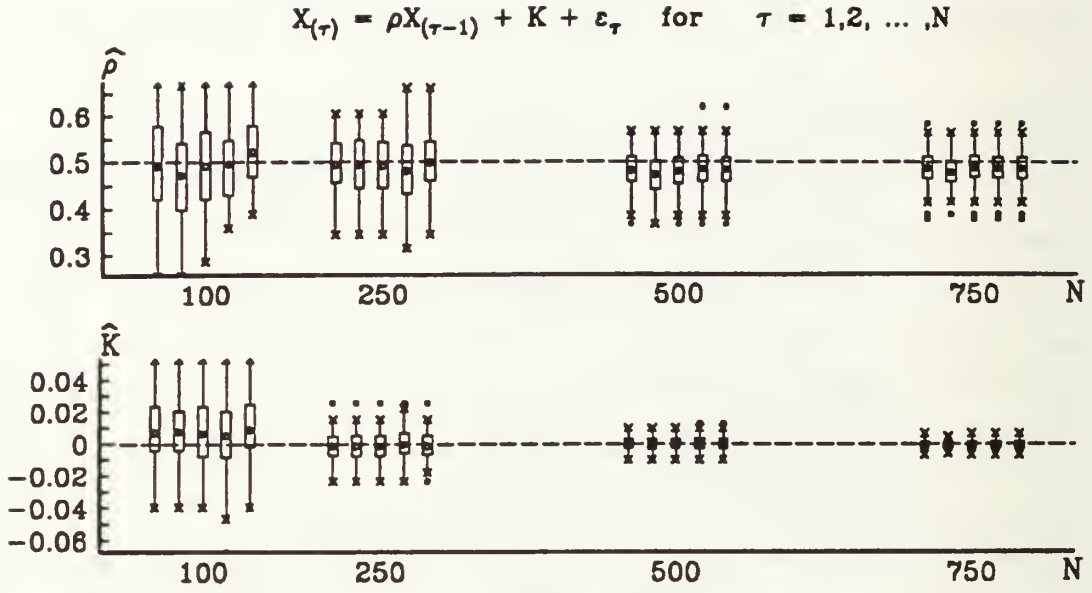

Figure 41. AR(1) MODEL SIMULATION: Boxplots of the estimates from each model selection criterion for $\rho = .5, K = 0$ when the model selection criterion within MARS correctly identified the data as from an AR(1) model (as reflected in Table 6). For increasing values of N, MARS attempted to identify the AR(1) model (64) from the simulated data using $P = 4$ lagged predictor variables and $M = 8$, the maximum number of subregions allowed in the forward-step MARS procedure, with $\sigma_\epsilon^2 = N(0,1)$ and a minimum span of MS = $.02(N)$ data points between knots. Each simulation consists of 100 replications. The model selection criterion represented by the boxplots are, from left to right and for each value of N; GCV^*, AIC, PC, SC and $AIC2$. The true value of the model coefficients, $\rho = .5$ and $K = 0$, are identified by the dashed line across each of the boxplots.

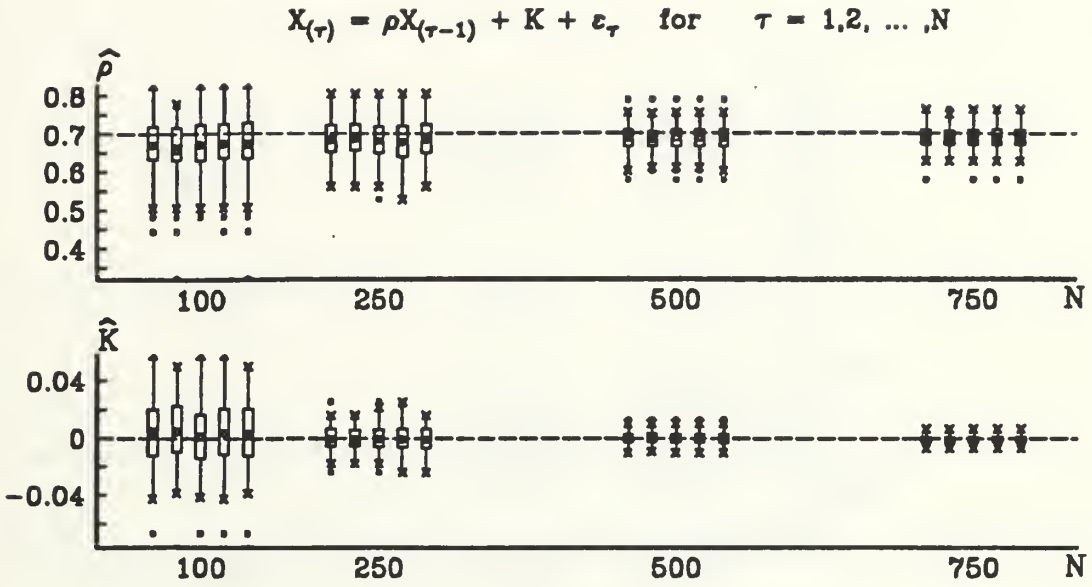

Figure 42. AR(1) MODEL SIMULATION: Boxplots of the estimates from each model selection criterion for $\rho = .7, K = 0$ when the model selection criterion within MARS correctly identified the data as from an AR(1) model (as reflected in Table 6). For increasing values of N, MARS attempted to identify the AR(1) model (64) from the simulated data using $P = 1$ lagged predictor variables and $M = 3$, the maximum number of subregions allowed in the forward-step MARS procedure, with $\sigma_\epsilon^2 = N(0,1)$ and a minimum span of MS = $.02(N)$ data points between knots. Each simulation consists of 100 replications. The model selection criterion represented by the boxplots are, from left to right and for each value of N; GCV^*, AIC, PC, SC and $AIC2$. The true value of the model coefficients, $\rho = .7$ and $K = 0$, are identified by the dashed line across each of the boxplots.

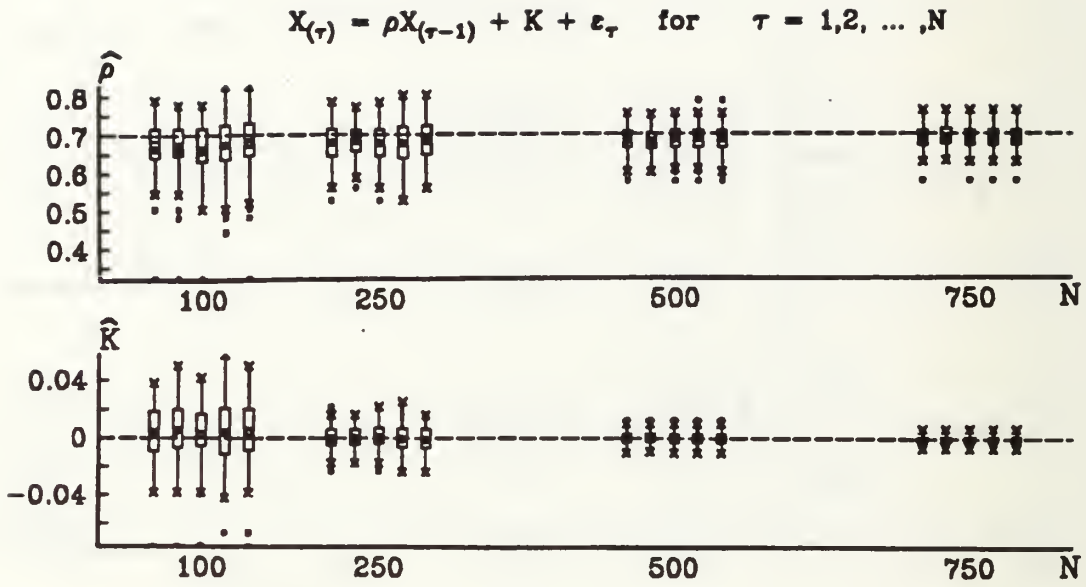

Figure 43. AR(1) MODEL SIMULATION: Boxplots of the estimates from each model selection criterion for $\rho = .7, K = 0$ when the model selection criterion within MARS correctly identified the data as from an AR(1) model (as reflected in Table 6). For increasing values of N, MARS attempted to identify the AR(1) model (64) from the simulated data using $P = 4$ lagged predictor variables and $M = 8$, the maximum number of subregions allowed in the forward-step MARS procedure, with $\sigma_\epsilon^2 = N(0,1)$ and a minimum span of MS = $.02(N)$ data points between knots. Each simulation consists of 100 replications. The model selection criterion represented by the boxplots are, from left to right and for each value of N; GCV^*, AIC, PC, SC and $AIC2$. The true value of the model coefficients, $\rho = .7$ and $K = 0$, are identified by the dashed line across each of the boxplots.

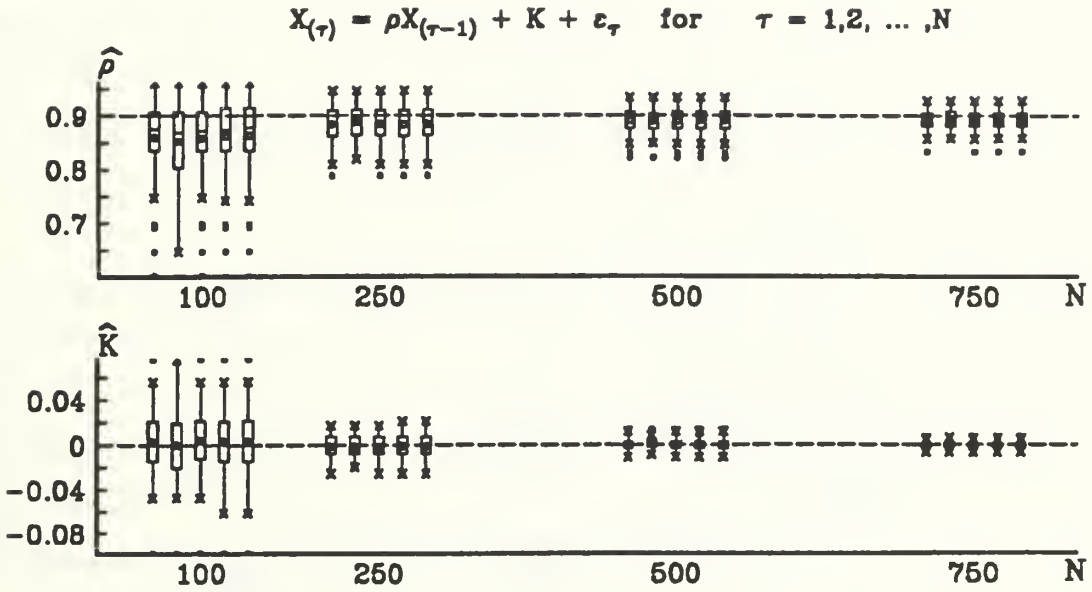

Figure 44. AR(1) MODEL SIMULATION: Boxplots of the estimates from each model selection criterion for $\rho = .9, K = 0$ when the model selection criterion within MARS correctly identified the data as from an AR(1) model (as reflected in Table 6). For increasing values of N, MARS attempted to identify the AR(1) model (64) from the simulated data using $P = 1$ lagged predictor variables and $M = 3$, the maximum number of subregions allowed in the forward-step MARS procedure, with $\sigma_\epsilon^2 = N(0,1)$ and a minimum span of MS = $.02(N)$ data points between knots. Each simulation consists of 100 replications. The model selection criterion represented by the boxplots are, from left to right and for each value of N; GCV^*, AIC, PC, SC and $AIC2$. The true value of the model coefficients, $\rho = .9$ and $K = 0$, are identified by the dashed line across each of the boxplots.

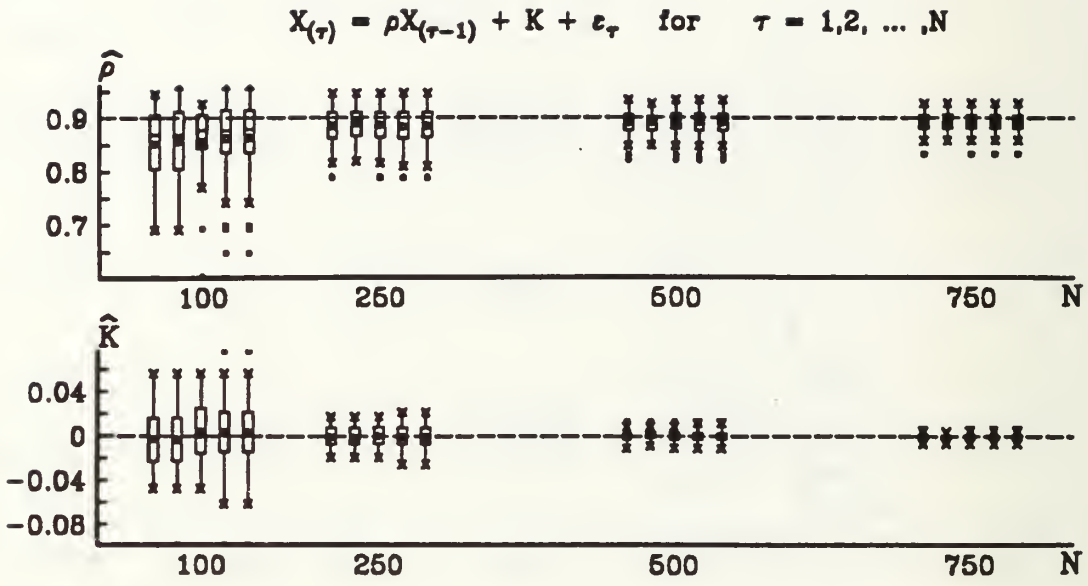

Figure 45. AR(1) MODEL SIMULATION: Boxplots of the estimates from each model selection criterion for $\rho = .9, K = 0$ when the model selection criterion within MARS correctly identified the data as from an AR(1) model (as reflected in Table 6). For increasing values of N, MARS attempted to identify the AR(1) model (64) from the simulated data using $P = 4$ lagged predictor variables and $M = 8$, the maximum number of subregions allowed in the forward-step MARS procedure, with $\sigma_\epsilon^2 = N(0,1)$ and a minimum span of MS = $.02(N)$ between knots. Each simulation consists of 100 replications. The model selection criterion represented by the boxplots are, from left to right and for each value of N; GCV^*, AIC, PC, SC and $AIC2$. The true value of the model coefficients, $\rho = .9$ and $K = 0$, are identified by the dashed line across each of the boxplots.

the assumed constant variance for ϵ_τ in both regions, X_τ can have a different variance in each of the two subregions. Also for a threshold at $t = 0$, the expected number of sample values in each subregion will be the same only if $\rho_1 = -\rho_2$.

Two categories of experiments were conducted using the threshold model.

The first experiment required each model selection criterion within MARS to estimate a model from the simulated data of the nonlinear threshold time series model using one lag predictor variable $X_{\tau-1}$, and using $M = 4$, the maximum number of subregions in the forward-step MARS procedure. The first experiment's alternative models include the constant model, linear autoregressive time series models, or nonlinear time series models that have more than one internal threshold.

The second experiment required each model selection criterion within MARS to estimate a model from the simulated data of the nonlinear threshold time series model where up to four lag predictor variables, $\{X_{\tau-i}\}_{i=1}^{4}$, are allowed, and using $M = 10$, the maximum number of subregions allowed in the forward-step MARS procedure. The second experiment's alternative models include the constant model, linear and nonlinear autoregressive time series models with terms other than $X_{\tau-1}$ (e.g. $X_{\tau-2}$), or nonlinear time series models with more than one internal threshold value on $X_{\tau-1}$.

Simulation experiments were performed for various combinations of ρ_1 and ρ_2 and for various values of the smoothing parameter MS, the minimum number of data points between knots on the same predictor variable. Table 7 and Figures 46–49 show the simulation results for $\rho_1, \rho_2 = .8, .4$, and $-.6, .6$, using $\sigma_\epsilon^2 = \mathrm{N}(0,.5)$ with the smoothing parameter MS $= .02(N)$ data points. Table 7 shows the number of simulations correctly identified as threshold time series models by each model selection criterion for a given length of the simulated time series N. On the left and right side of the table are the results of the first and second experiments in which MARS attempted to identify the nonlinear threshold time series model (65) from the simulated data using $P = 1$ (left) and $P = 4$ (right) lagged predictor variables and using $M = 4$ (left) and $M = 10$ (right), the

Overall, the SC and $AIC2$ criteria perform the best at correctly identifying the simulated data as the simple nonlinear threshold time series model (65). For the first experiment (left), $P = 1$ and $M = 4$, all the model selection criteria appear to perform equally well. In the second experiment (right), $P = 4$ and $M = 10$, the SC and $AIC2$

TABLE 7. THRESHOLD MODEL SIMULATION: The number of threshold simulations correctly identified by each model selection criterion within MARS for increasing values of N. The model parameters are $\rho_1, \rho_2 = .8, .4$ (top) and $-.6, .6$ (bottom), and $t = 0$, with $\sigma_\epsilon^2 = N(0, .5)$ and a minimum span of MS $= .02(N)$ data points between knots. MARS attempted to identify the AR(1) model (64) from the simulated data using $P = 1$ (left) and $P = 4$ (right) lagged predictor variables and using $M = 4$ (left) and $M = 10$ (right), the maximum number of subregions allowed in the forward-step MARS procedure. Each simulation consists of 100 replications. All the model selection criterion perform well in identifying the threshold simulations.

	$P = 1$ and $M = 4$				$P = 4$ and $M = 10$			
N	500	750	1000	1500	500	750	1000	1500
	$\rho_1, \rho_2 = .8, .4$							
*GCV**	54	89	96	99	52	75	89	97
AIC	77	94	93	96	60	71	72	79
PC	57	88	97	98	48	77	89	95
SC	61	78	89	97	61	78	89	97
AIC2	66	84	93	97	66	84	93	97
	$\rho_1, \rho_2 = -.6, .6$							
*GCV**	94	100	100	100	77	83	89	94
AIC	95	99	97	95	77	76	74	76
PC	93	100	100	100	78	80	87	92
SC	94	100	100	100	94	100	100	100
AIC2	98	100	100	100	98	99	99	99

criteria perform the best followed closely by the PC and GCV^* criteria as the values of N increase. However, the AIC criterion's performance in the second experiment does not improve for increasing values of N and falls behind at correctly identifying the nonlinear threshold time series model (65). Again, as in the experiments with the AR(1) simulations, the majority of models incorrectly identified by the AIC criterion included additional terms i.e., AIC appears to overestimate the number of parameters in the model.

Figures 46–49 are a series of box plots for the estimated coefficients of the simulations correctly identified as threshold time series models (as addressed in Table 7) by each model selection criterion within MARS for increasing values of N. For each value of N, the boxplots represent the estimated model coefficients using, from left to right, GCV^*, AIC, PC, SC and $AIC2$. The estimates for $\hat{\rho}_1$ (top), $\hat{\rho}_2$ (middle) and $\hat{t}$ (bottom) are given. The true value of each model coefficient, $\rho_1, \rho_2 = .8, .4$ (Figures 46–47), $\rho_1, \rho_2 = -.6, .6$ (Figures 48–49), and $t = 0$, are identified by the dashed line across each of the box plots. At the bottom of each boxplot is the length N of each simulated time series. It is observed that the estimated values of the model coefficients tend to their true value as N increases. Due to several outliers, the performance of PC and GCV^* at estimating ρ_2 are initially disappointing.

3. Summary of AR(1) and Threshold Model Simulations

Overall, SC was the best criterion at identifying and selecting the model coefficients from the simulated data of these simple AR(1) and nonlinear threshold time series models. The SC criterion was consistent for all values of the AR(1) and threshold model coefficients while the other criterion at times had difficulty especially for small N (see e.g. Table 6 with $\rho = .5$). The performance of the SC criterion was followed next by the $AIC2$ criterion and then the PC and GCV^* criteria. In all cases with the exception of AIC, the number of correctly identified simulation models improved for increasing values of N. Also for increasing values of N, when the AR(1) or nonlinear threshold time series model was correctly identified, the precision of the estimates of the model coefficients improved for each model selection criterion. It was noted that when AIC incorrectly identified a model, it added additional terms to the model (in almost all cases), i.e., AIC was able to

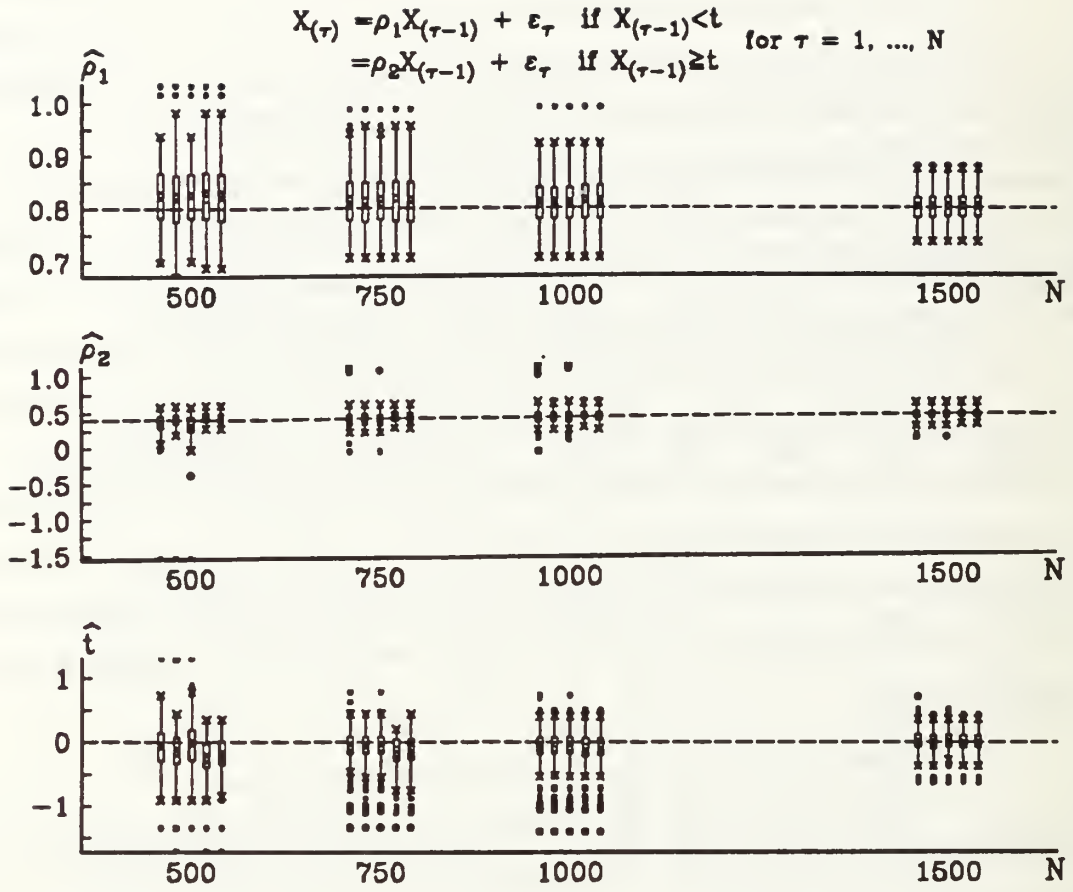

Figure 46. THRESHOLD MODEL SIMULATION: Boxplots of the estimates for $\rho_1, \rho_2 = .8.4$, and $t = 0$ when the model selection criterion within MARS correctly identified the data as a threshold model (as reflected in Table 7). For increasing values of N, MARS attempted to identify the threshold model (33) from the simulated data using $P = 1$ lagged predictor variables and $M = 4$, the maximum number of subregions allowed in the forward-step MARS procedure, with $\sigma_\epsilon^2 = N(0, .5)$ and a minimum span of MS $= .02(N)$ data points between knots. Each simulation consists of 100 replications. The model selection criterion represented by the boxplots are, from left to right and for each value of N; GCV^*, AIC, PC, SC and $AIC2$. The true value of the model coefficients, $\rho_1 = .8, \rho_2 = .4$ and $t = 0$, are identified by the dashed line across each of the boxplots.

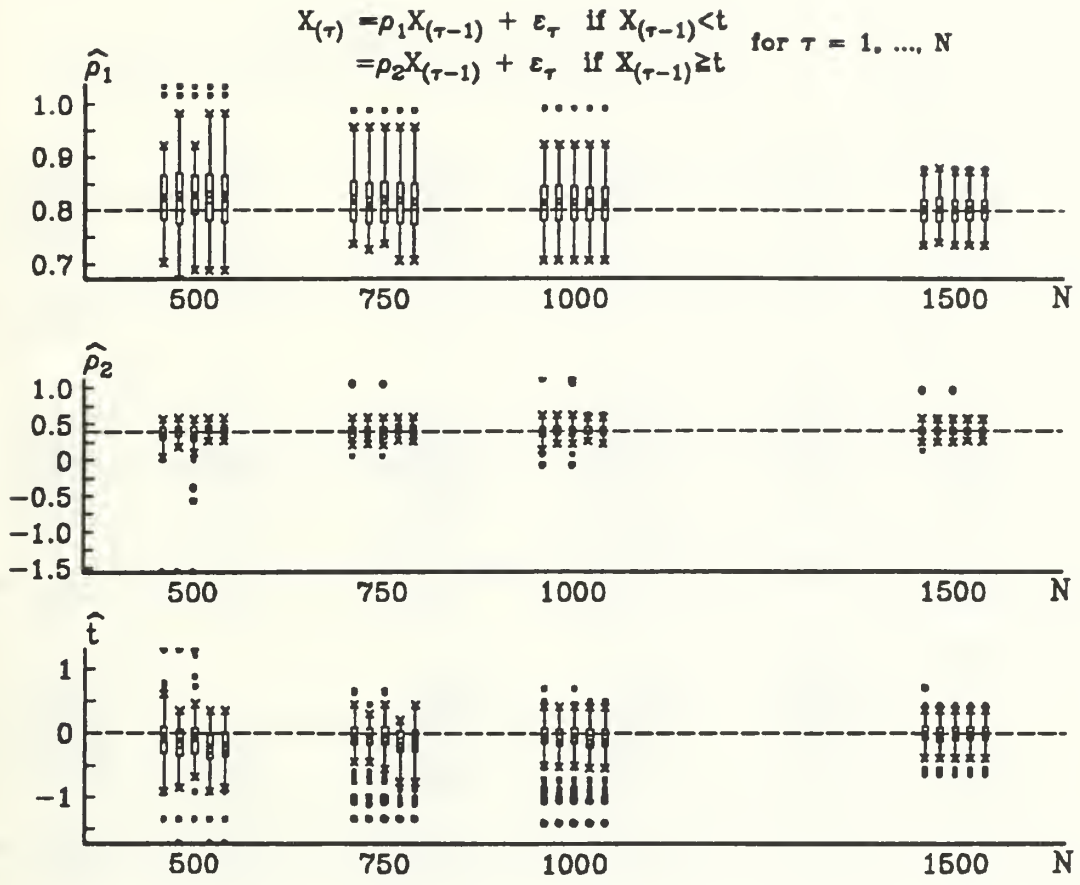

Figure 47. THRESHOLD MODEL SIMULATION: Boxplots of the estimates for $\rho_1, \rho_2 = .8.4$, and $t = 0$ when the model selection criterion within MARS correctly identified the data as a threshold model (as reflected in Table 7). For increasing values of N, MARS attempted to identify the threshold model (33) from the simulated data using $P = 4$ lagged predictor variables and $M = 10$, the maximum number of subregions allowed in the forward-step MARS procedure, with $\sigma_\epsilon^2 = N(0, .5)$ and a minimum span of MS $= .02(N)$ data points between knots. Each simulation consists of 100 replications. The model selection criterion represented by the boxplots are, from left to right and for each value of N; GCV^*, AIC, PC, SC and $AIC2$. The true value of the model coefficients, $\rho_1 = .8, \rho_2 = .4$ and $t = 0$, are identified by the dashed line across each of the boxplots.

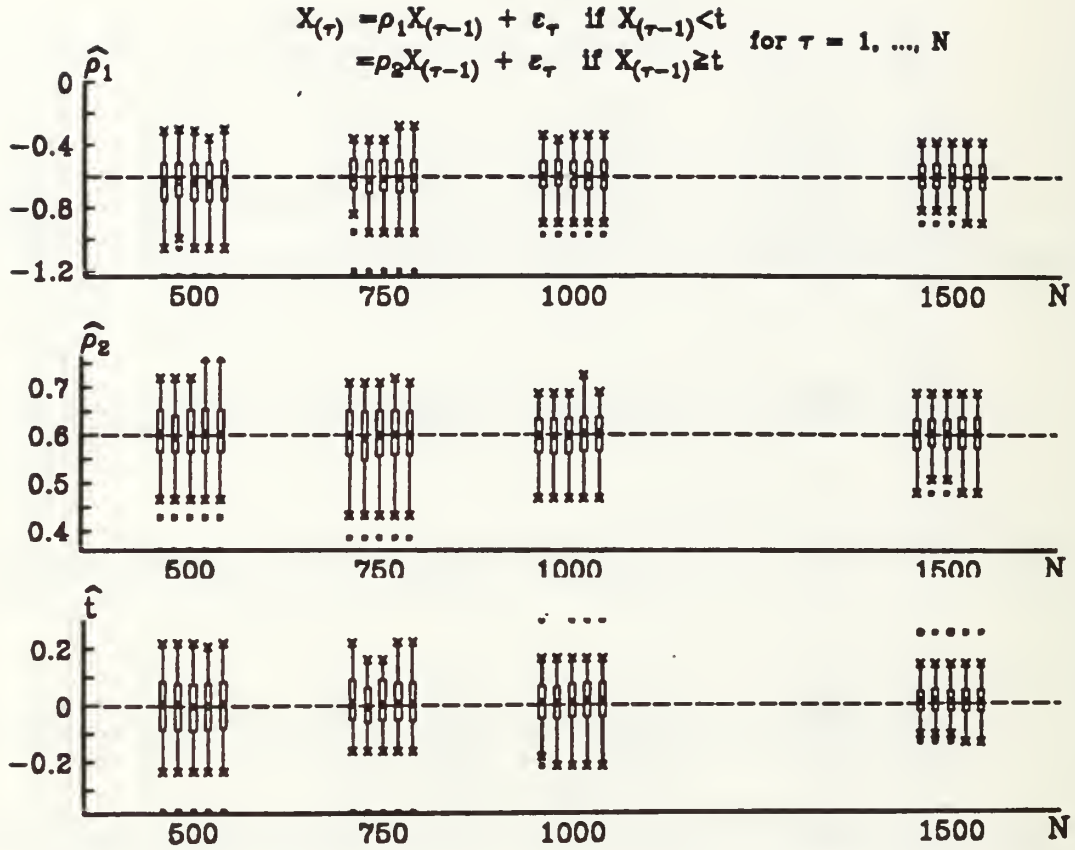

Figure 48. THRESHOLD MODEL SIMULATION: Boxplots of the estimates for $\rho_1, \rho_2 = -.6.6$, and $t = 0$ when the model selection criterion within MARS correctly identified the data as a threshold model (as reflected in Table 7). For increasing values of N, MARS attempted to identify the threshold model (65) from the simulated data using $P = 1$ lagged predictor variables and $M = 4$, the maximum number of subregions allowed in the forward-step MARS procedure, with $\sigma_\epsilon^2 = N(0, .5)$ and a minimum span of MS $= .02(N)$ data points between knots. Each simulation consists of 100 replications. The model selection criterion represented by the boxplots are, from left to right and for each value of N; GCV^*, AIC, PC, SC and $AIC2$. The true value of the model coefficients, $\rho_1 = -.6, \rho_2 = .6$ and $t = 0$, are identified by the dashed line across each of the boxplots.

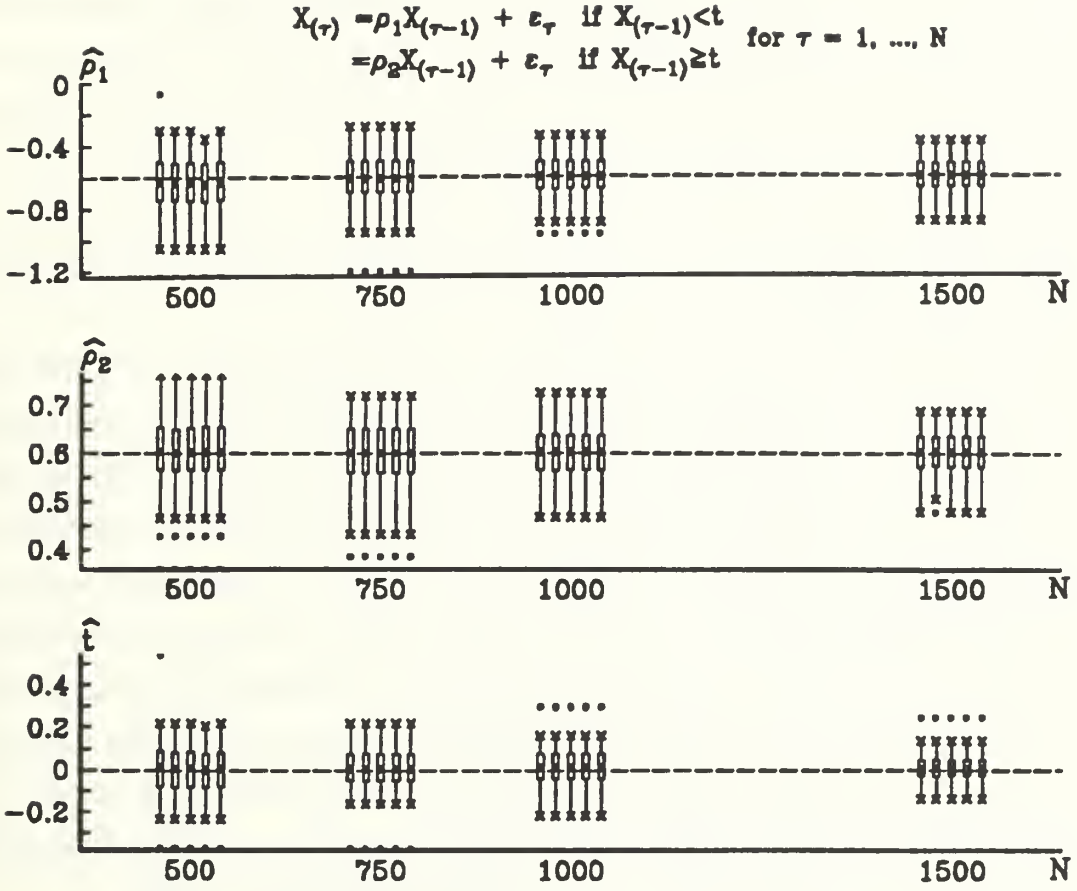

Figure 49. THRESHOLD MODEL SIMULATION: Boxplots of the estimates for $\rho_1, \rho_2 = -.6.6$, and $t = 0$ when the model selection criterion within MARS correctly identified the data as a threshold model (as reflected in Table 7). For increasing values of N, MARS attempted to identify the threshold model (65) from the simulated data using $P = 4$ lagged predictor variables and $M = 10$, the maximum number of subregions allowed in the forward-step MARS procedure, with $\sigma_\epsilon^2 = N(0, .5)$ and a minimum span of MS $= .02(N)$ data points between knots. Each simulation consists of 100 replications. The model selection criterion represented by the boxplots are, from left to right and for each value of N; GCV^*, AIC, PC, SC and $AIC2$. The true value of the model coefficients, $\rho_1 = -.6, \rho_2 = .6$ and $t = 0$, are identified by the dashed line across each of the boxplots.

identify structure, although more than was actually present, which agrees with work done by Schwarz (1978).

Another approach for determining the relative ability of each model selection criterion is to evaluate their performance at approximating the response variable (input) in terms of the explanatory variables (output). In this regard, the next section investigates the ability of each model selection criterion to approximate the fitted values and limit cycle from ASTAR Model 9 of the Wolf sunspot numbers.

D. SIMULATIONS OF ASTAR MODEL 9 OF THE WOLF SUNSPOT NUMBERS

As an illustration of the relative ability of each model selection criterion within MARS to closely approximate a representation of an actual time series we used the fitted values and limit cycle of ASTAR Model 9 of the Wolf sunspot numbers (66). In the first part of this section the Wolf sunspot numbers and the fitted values and limit cycle from ASTAR Model 9 are briefly reviewed. (Recall that Chapter II discussed the use of MARS for modeling and prediction of the Wolf sunspot numbers, an actual time series with periodic behavior. The result was ASTAR Model 9, which when used for prediction was a considerable improvement over previous existing nonlinear models of the Wolf sunspot numbers.) Next, two simulations are used to examine the ability of each model selection criterion to closely approximate the fitted value and limit cycle time series of ASTAR Model 9 from the lagged values of each respective time series with additive N(0,1) noise.

The sunspot data and the fitted values of ASTAR Model 9 (Figure 50) are quite 'periodic' but have nonsymmetric cycles with extremely sharp peaks and troughs. The cycles generally vary between 10 and 12 years with the greater number of sunspots concentrated in each descent period versus the accompanying ascent period. The average (ascent/descent) period is (4.6/6.6) years for the sunspot number data and (4.5/6.4) years for the fitted values from ASTAR Model 9. The functional form of ASTAR Model 9 is

$$X_\tau = \begin{cases} 2.711 + .960X_{\tau-1} + .332(47.0 - X_{\tau-5})_+ - .257(59.1 - X_{\tau-9})_+ \\ - .003X_{\tau-1}(X_{\tau-2} - 26.0)_+ + .017X_{\tau-1}(44.0 - X_{\tau-3})_+ \\ - .032X_{\tau-1}(17.1 - X_{\tau-4})_+ \\ + .004X_{\tau-1}(26.0 - X_{\tau-2})_+(X_{\tau-5} - 41.0)_+ + \epsilon_\tau \end{cases} \tag{66}$$

where $(x)_+$ is a plus function i.e., a function which takes value x if $x > 0$ and takes values 0 otherwise, and ϵ_τ (from the analysis in Chapter II) is assumed to be Gaussian noise with zero mean and variance σ_ϵ^2. Model 9 has 14 parameters with 8 terms (a constant term with 3 one-way, 3 two-way and 1 three-way interactions) and 6 threshold values (1 each on $X_{\tau-2}$, $X_{\tau-3}$, $X_{\tau-4}$, and $X_{\tau-9}$ and 2 on $X_{\tau-5}$). Note that the MARS algorithm generating ASTAR Model 9 uses 20 lagged predictor variables that are permitted to form 1, 2, and 3-way interactions during a maximum of $M = 15$ steps of the forward-step MARS algorithm. The minimum span between threshold knots is MS = 18 data points.

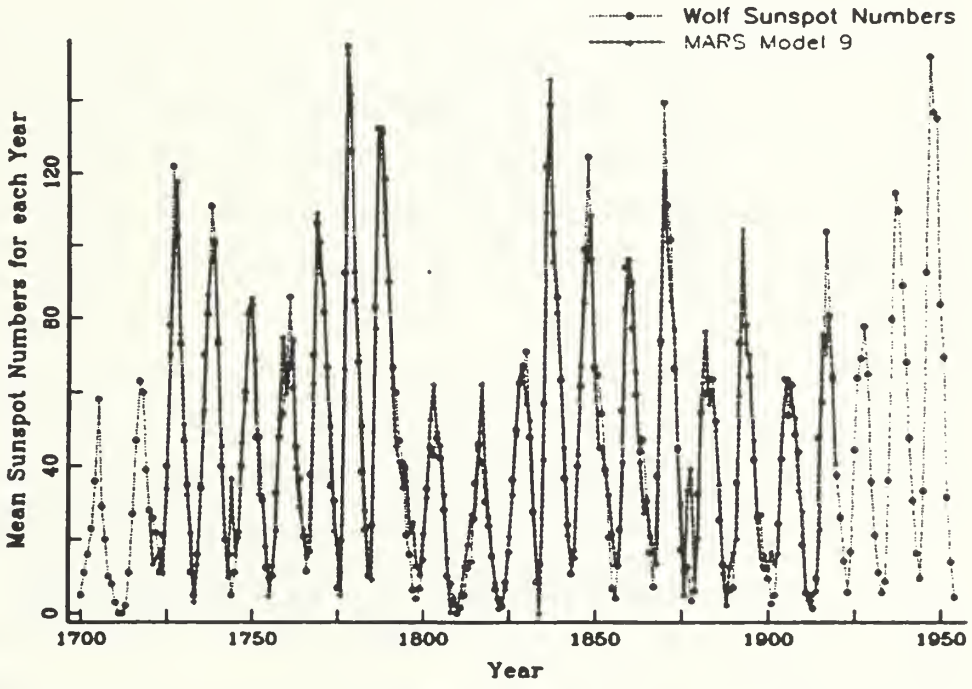

Figure 50. The yearly Wolf sunspot numbers (1700-1955) versus the fit of ASTAR Model 9 (1720-1920). The yearly sunspot numbers (1700-1719) were used for initialization. The yearly sunspot numbers (1921-1955) were used to examine the prediction performance of ASTAR Model 9 and other models of the yearly sunspot numbers.

One of the interesting aspects of ASTAR models is their ability to create models with limit cycles from periodic-like data such as the Wolf sunspot numbers. A limit cycle may be thought of as a stationary state of sustained oscillations (Tong, 1985). Figure 51 shows the 137 year limit cycle of ASTAR Model 9 of the Wolf sunspot numbers with its ascent and descent periods. The limit cycle for Model 9 is asymmetric with a range in amplitude of 17.7 to 94.5 and an average ascent/descent period of 4.3/6.3 years versus the 4.6/6.6 years for the actual yearly sunspot numbers from 1700 to 1920. In comparing Model 9's limit cycle (Figure 51) with the real yearly sunspot data (Figure 50) note that the standard deviation of the fitted residual's error variance is estimated as $(MSS)^{1/2} = 10.69$ sunspots.

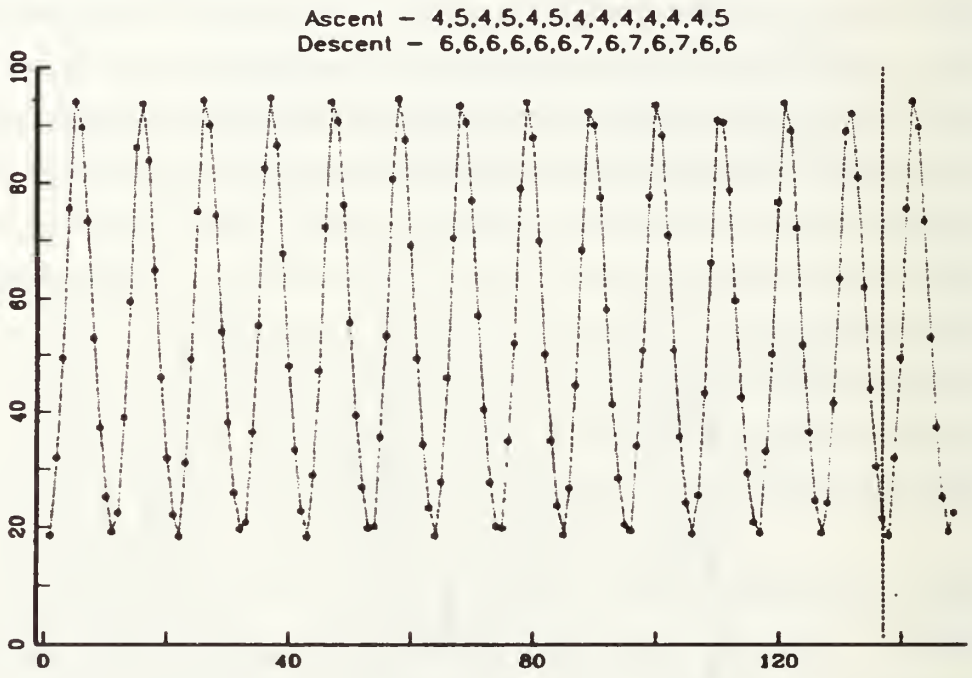

Figure 51. The limit cycle for ASTAR Model 9 of the yearly Wolf sunspot numbers (1720-1920). The limit cycle is 137 years long with the indicated ascent and descent periods. The limit cycle is generated using ASTAR Model 9 initialized with the yearly sunspot numbers (1700-1719). The 'subcycles' have lengths of 10 or 11 years with 4 or 5 years per ascent period and 6 or 7 years per descent period.

1. Simulations of ASTAR Model 9

Two different simulations using ASTAR Model 9 were developed to evaluate each model selection criterion within MARS. The first simulation experiment used the fitted

values of Model 9 (Figure 50) plus additive N(0,σ_ϵ^2) noise to examine the ability of each criterion to model and estimate a time series with rapidly changing structure. The second experiment used the limit cycle produced by Model 9 (Figure 51) plus additive N(0,σ_ϵ^2) noise to examine the ability of each criterion to model and estimate a time series with repetitive structure. The objective of these simulations are two fold; to observe how well each model selection criterion estimates the fitted values and the limit cycle of ASTAR Model 9 and how consistent are these estimates. Mean square error (MSE) was used as an overall measure of performance of each model selection criterion.

a. Simulating the Fitted Values of ASTAR Model 9

In this first experiment simulations of the fitted values of ASTAR Model 9 of the Wolf sunspot numbers were considered. Using the fitted values of ASTAR Model 9 to represent $f(\boldsymbol{X})$, independent N(0,1) noise was added for the 221 year period from 1700 to 1920. The resulting values were used as the time series for input to the MARS program. The program parameters in MARS during each simulation remained the same as those used to develop ASTAR Model 9; $P = 20$ lagged predictor variables, a maximum level of MI = 3 interactions i.e., the models were permitted to form 1, 2, and 3-way interactions, and a minimum span between knots on a lagged predictor variable of MS = 18 data points. The data values from 1700 to 1719 were used for model initialization. A total of 50 simulations were performed for each different model selection criterion for different values of M that ranged from 5 to 30 (M is the maximum number of forward steps in the MARS algorithm). Note that a low value of M (e.g. 5 or 10) does not adequately permit a model selection criterion in MARS to find the structure of Model 9's fitted values during the forward-step algorithm. In contrast, a high value of M (e.g. 25 or 30) does permit the model selection criterion to find the structure of Model 9's fitted values and requires the backward-step algorithm to trim excess terms from the model.

Table 8 and Figures 52 and 53 show examples of the results from the simulations of the fitted values of ASTAR Model 9. For each value of M and each model selection criterion, the bias and variance for each estimate of the 201 (1720-1920) actual fitted sunspot numbers from ASTAR Model 9 was computed, using the results from the 50 simulation runs. The values in Table 8 represent the average across $\tau = 1, \ldots, 201$ for the absolute bias, variance and MSE of the estimates of the fitted values from ASTAR Model 9

using each model selection criterion. As expected, bias dominated the MSE due to the rapidly changing structure in the fitted values from Model 9. Also, as the models in MARS are permitted to become more complex (M increases) the bias decreases and, in general, the variance of the estimates of the fitted values from Model 9 increase.

Figures 52 and 53 are plots of the bias [points] and the range of a 95% confidence interval centered about zero ($\pm 1.64\sigma/N^{.5}$) [lines] for each of the 201 estimates of the fitted values of Model 9 from the 50 simulations for each identified model selection criterion. In Figure 52 are the results for the model selection criteria AIC (left column) and GCV^* (right column) with $M = 10$ (top) and $M = 25$ (bottom). In Figure 53 are the results of the model selection criteria for AIC (left column) and PC (right column) with $M = 10$ (top) and $M = 25$ (bottom). Note the difference in the size of the bias and the size of the confidence interval between the values of M in the top and bottom plots of each Figure, the bias for each estimate being, in general, smaller for $M = 25$ while the size of the confidence interval generally increases. Using Table 8 the AIC criterion is better than GCV^* for both $M = 10$ and $M = 25$ while AIC is better than PC for $M = 10$ but they appear equivalent for $M = 30$. Looking between the plots of AIC and GCV^* (Figure 52) for $M = 10$ (top plots) note the high positive bias in several estimates of the fitted values using GCV^* while for $M = 25$ (bottom plots) note the spread of the CI using GCV^*. Looking between the plots of AIC and PC (Figure 53) for $M = 10$ (top plots) note the high positive bias in several estimates of the fitted values using PC. By looking across and down in Figures 52 and 53 and using Table 8 it is observed that the MSE of each model selection criterion is improving for increasing values of M although the rate of improvement decreases as M increases.

Using the simulation results from Table 8 and plots like those in Figures 52 and 53, AIC is the best model selection criterion for estimating the fitted values of ASTAR Model 9 using MSE as the measure of performance. In Table 8, for each value of M the average absolute bias and average MSE across τ for the AIC criterion is, in general, lower than the other model selection criteria. Recall that AIC tends to over-parameterize a model, which may explain AIC's performance for this experiment. The AIC criterion's performance is closely followed by the PC, SC, and GCV^* criteria. $AIC2$'s performance is extremely poor throughout the experiment.

TABLE 8. SIMULATION of the FITTED VALUES of ASTAR MODEL 9: The average across $\tau = 1, \ldots, 201$ of the absolute bias, variance and MSE of the estimates for the fitted values of ASTAR Model 9 from each model selection criterion within MARS using 50 simulations for increasing values of M, the maximum number of forward-step subregions permitted in a MARS model. The MARS parameters for each of the 50 simulations and each model selection criterion are $P = 20$, and $\sigma_\epsilon^2 = N(0,1)$ with a minimum span of MS = 18 data points between threshold knots. Each simulation consisted of estimating the 201 fitted values from ASTAR MODEL 9 of the Wolf sunspot numbers with additive N(0,1) noise.

Average Absolute Bias

M	5	10	15	20	25	30
GCV^*	12.07	9.96	8.93	7.84	7.23	6.78
AIC	11.53	9.66	8.40	7.61	6.98	6.48
PC	12.05	9.96	8.84	7.73	7.14	6.49
SC	11.59	9.92	8.72	8.01	7.55	7.34
$AIC2$	11.94	10.81	11.56	11.54	11.66	11.77

Average Variance

GCV^*	0.408	0.750	0.949	1.040	1.043	1.024
AIC	0.428	0.838	0.849	0.863	0.875	0.884
PC	0.399	0.750	0.942	1.028	0.988	0.945
SC	0.423	0.750	0.905	0.922	1.002	0.991
$AIC2$	0.480	0.687	0.929	1.005	0.978	0.967

AVERAGE MSE

GCV^*	245.5	173.5	132.7	104.0	87.6	75.5
AIC	235.1	158.4	117.8	96.8	82.0	70.9
PC	244.8	173.5	130.7	101.1	84.6	69.7
SC	237.3	172.6	126.4	107.6	95.4	89.9
$AIC2$	251.8	239.0	223.0	220.6	223.9	228.5

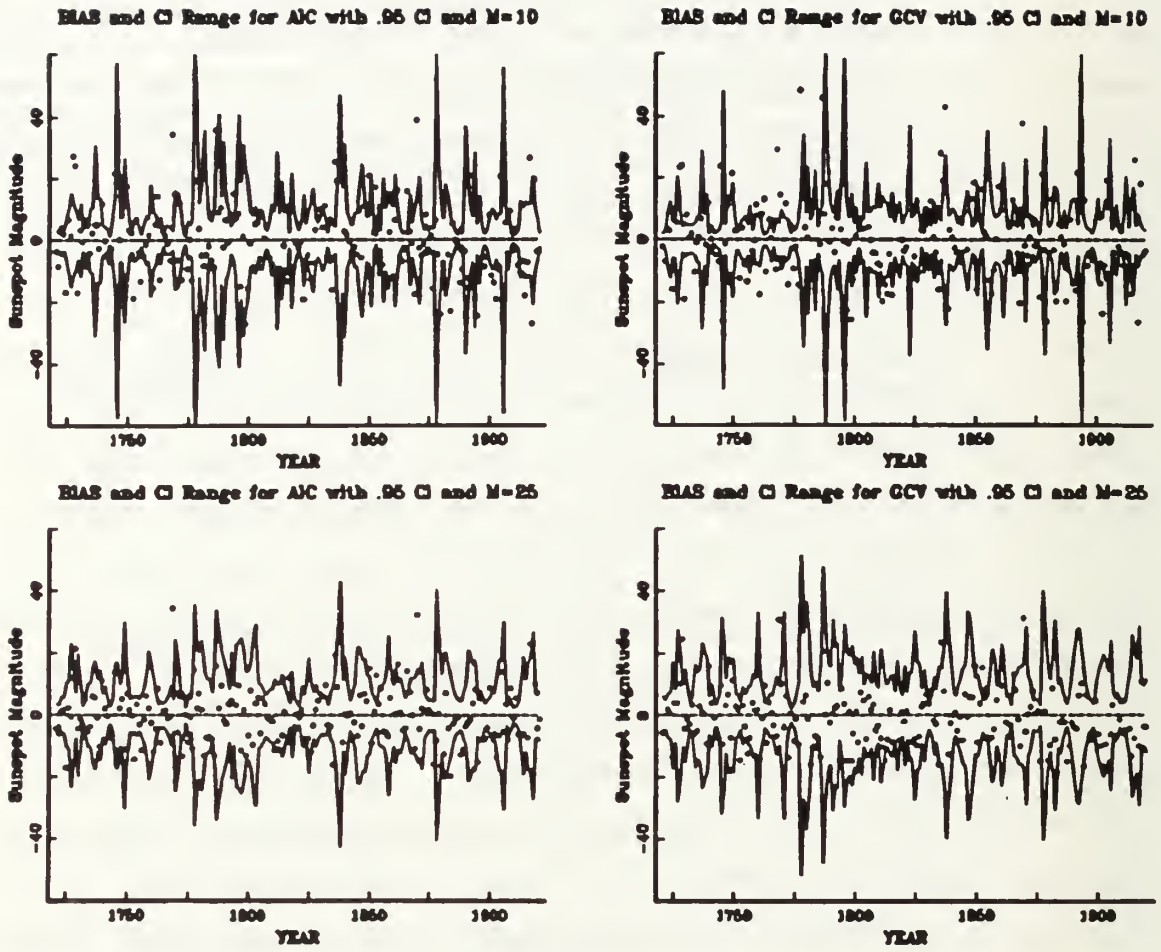

Figure 52. SIMULATION of the FITTED VALUES of ASTAR MODEL 9: The bias (points) and a 95% confidence interval centered about zero (lines) for the estimates of the fitted values of ASTAR Model 9. The simulation experiment used 50 simulations of the 221 fitted values from ASTAR Model 9 of the Wolf Sunspot numbers with additive N(0,1) noise. The plots in this figure are for the AIC [left] and GCV^* [right] model selection criteria using $M = 10$ [top] and $M = 25$ [bottom], the maximum number of subregions permitted in the forward step of the MARS algorithm.

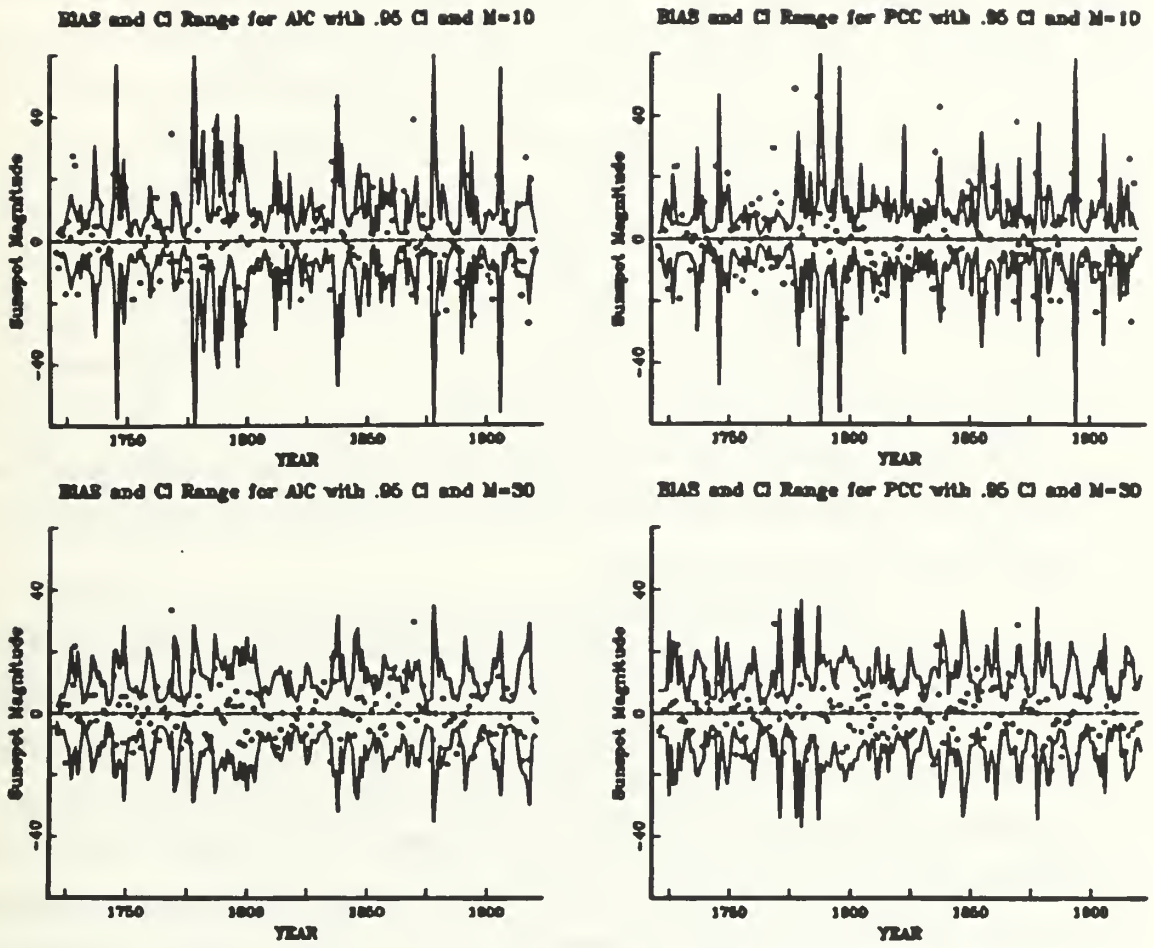

Figure 53. SIMULATION of the FITTED VALUES of ASTAR MODEL 9: The bias (points) and a 95% confidence interval centered about zero (lines) for the estimates of the fitted values of ASTAR Model 9. The simulation experiment used 50 simulations of the 221 fitted values from ASTAR Model 9 of the Wolf Sunspot numbers with additive N(0,1) noise. The plots in this figure are for the *AIC* [left] and *PC* [right] model selection criteria using $M = 10$ [top] and $M = 30$ [bottom], the maximum number of subregions permitted in the forward step of the MARS algorithm.

b. *Simulations of the Limit Cycle from ASTAR Model 9*

In the second experiment we considered simulations of the limit cycle developed using ASTAR Model 9 of the Wolf sunspot numbers. Using (66) and the Sunspot numbers from 1700-1719 for initialization, Model 9's limit cycle is created by driving the model without noise. The resulting 137 year limit cycle is shown at Figure 51. Using the limit cycle of ASTAR Model 9 to represent $f(X)$, independent N(0,1) noise was added to 431 limit cycle data values (3 cycles of 137 values and 20 values for initialization). The resulting values were used as the time series for input into the MARS program. The program parameters in MARS during each simulation remained the same as those used to develop ASTAR Model 9; $P = 20$ lagged predictor variables, a maximum level of MI = 3 interactions, i.e., the models were permitted to form 1, 2, and 3-way interactions, and a minimum span between variable partitions of MS = 18 data points. A total of 50 simulations were performed for each model selection criterion for different values of M that range from 5 to 15. Fewer number of forward steps M are required in this experiment due to the repetitiveness of Model 9's limit cycle.

Table 9 and Figures 54 and 55 show examples of the results from the second simulation experiment. For each value of M and each model selection criterion, the bias and variance for each estimate of the 411 limit cycle values was computed, using the results from the 50 simulation runs. The values in Table 9 represent the average across $\tau = 1, \ldots, 411$ of the absolute bias, variance and MSE of the estimates of the limit cycle values from ASTAR Model 9 using each model selection criterion. The bias again dominates the MSE although it is not as significant as the bias in the estimates of the fitted values from ASTAR Model 9 (Table 8). Note again that as the models in MARS are permitted to become more complex (M increases) the bias decreases. Also, the variance of the estimates for the limit cycle values are, in general, slowly decreasing.

Figures 54 and 55 are plots of the bias [points] and the range of a 95% confidence interval centered about zero ($\pm 1.64\sigma/N^{.5}$) [lines] for each of the 411 estimates of ASTAR Model 9's limit cycle using the 50 simulations and the identified model selection criterion. Figure 54 shows the results for AIC (left column) and GCV^* (right column) with $M = 5$ (top) and $M = 15$ (bottom). Figure 55 shows the results for PC (left column) and GCV^* (right column) with $M = 10$ (top) and $M = 30$ (bottom). Again, note the difference

TABLE 9. SIMULATION of the LIMIT CYCLE VALUES of ASTAR MODEL 9: The average across $\tau = 1, \ldots, 411$ of the absolute bias, variance and MSE of the estimates for the limit cycles values from ASTAR Model 9 for each model selection criterion within MARS using 50 simulations and increasing values of M, the maximum number of forward-step subregions permitted in a MARS model. The MARS parameters for each of the 50 simulation are $P = 20$, $\sigma_\epsilon^2 = N(0,1)$, with a minimum span of MS = 18 data points between model threshold knots. Each simulation consisted of estimating 411 values of ASTAR Model 9's limit cycle with additive N(0,1) noise (3 limit cycles of 137 data values).

	Average Absolute Bias			Average Variance			Average MSE		
M	5	10	15	5	10	15	5	10	15
GCV^*	2.42	1.73	1.46	0.12	0.11	0.10	9.41	4.98	3.59
AIC	2.21	1.61	1.44	0.14	0.10	0.09	8.10	4.32	3.48
PC	2.42	1.74	1.47	0.12	0.11	0.10	9.41	5.00	3.63
SC	2.21	1.63	1.53	0.14	0.10	0.10	8.03	4.45	3.87
$AIC2$	2.21	1.94	1.82	0.14	0.13	0.14	7.95	5.83	5.21

in the size of the bias and the size of the confidence interval between the values of M in the top and bottom plots of each Figure; the bias and the variance for each estimate being, in general, smaller for $M = 15$. By looking across and down in Figures 54 and 55 and using Table 9 it is again observed that the MSE of each model selection criterion is improving for increasing values of M although again the rate of improvement decreases as M increases.

Using the simulation results from Table 9 and plots like those in Figures 54 and 55, AIC is the best model selection criterion for estimating the limit cycle values of ASTAR Model 9 using MSE as the measure of performance. The AIC criterion's performance is closely followed by the SC, PC and GCV^*. $AIC2$'s performance, initially good at $M = 5$, is again poor for increasing values of M.

c. *Summary of ASTAR Model 9 Simulations*

The AIC criterion performed very well for the simulations of the fitted values and the limit cycle of ASTAR Model 9. The performance of the AIC criterion was followed closely by the SC, PC and GCV^* criteria, with SC initially doing better for lower values of M. Overall, the $AIC2$ criterion performed poorly throughout the experiment.

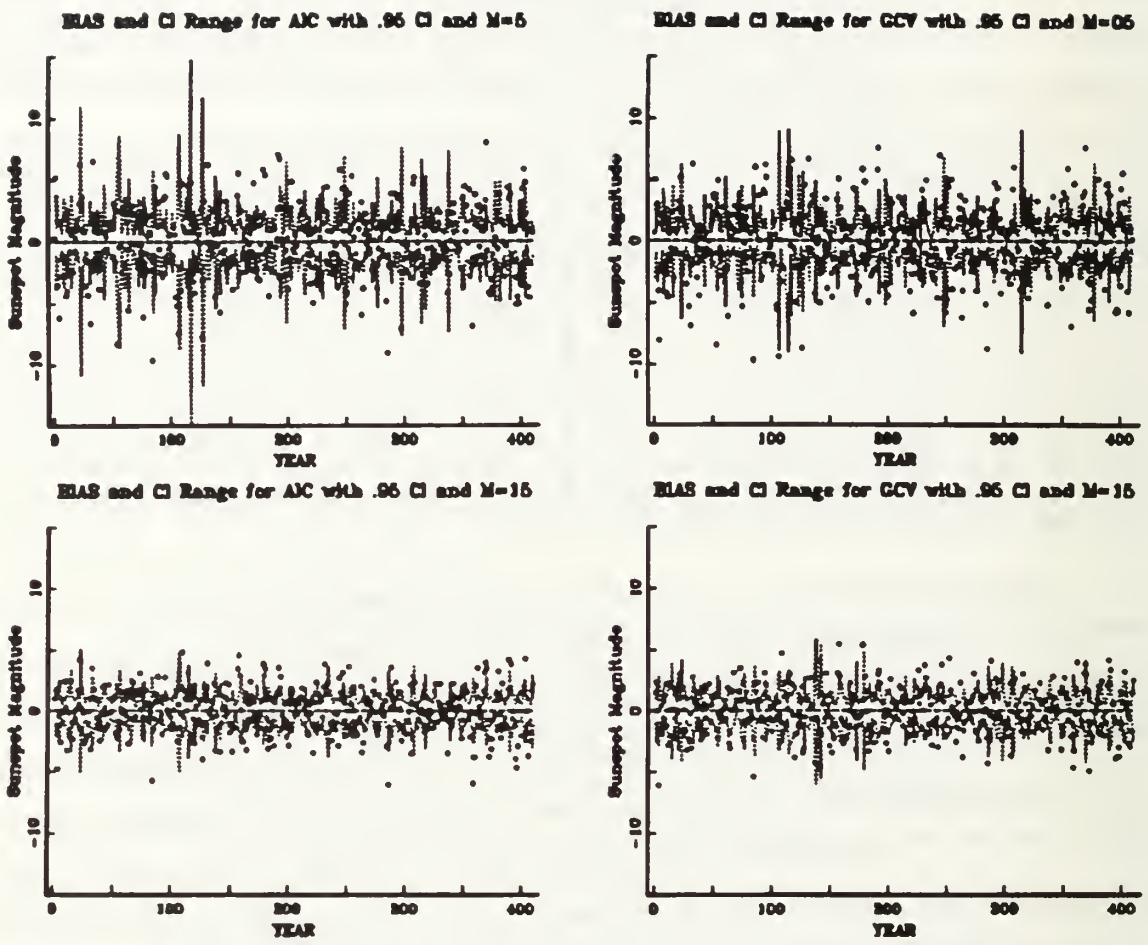

Figure 54. SIMULATION of the LIMIT CYCLE from ASTAR MODEL 9: The bias (points) and a 95% confidence interval centered about zero (lines) for the estimates of ASTAR Model 9's limit cycle. The simulation experiment used 50 simulations of 411 values of the limit cycle developed from ASTAR Model 9 of the Wolf Sunspot numbers with additive N(0,1) noise. The plots in this figure are for the AIC [left] and GCV^* [right] model selection criteria using $M = 5$ [top] and $M = 15$ [bottom], the maximum number of subregions permitted in the forward step of the MARS algorithm.

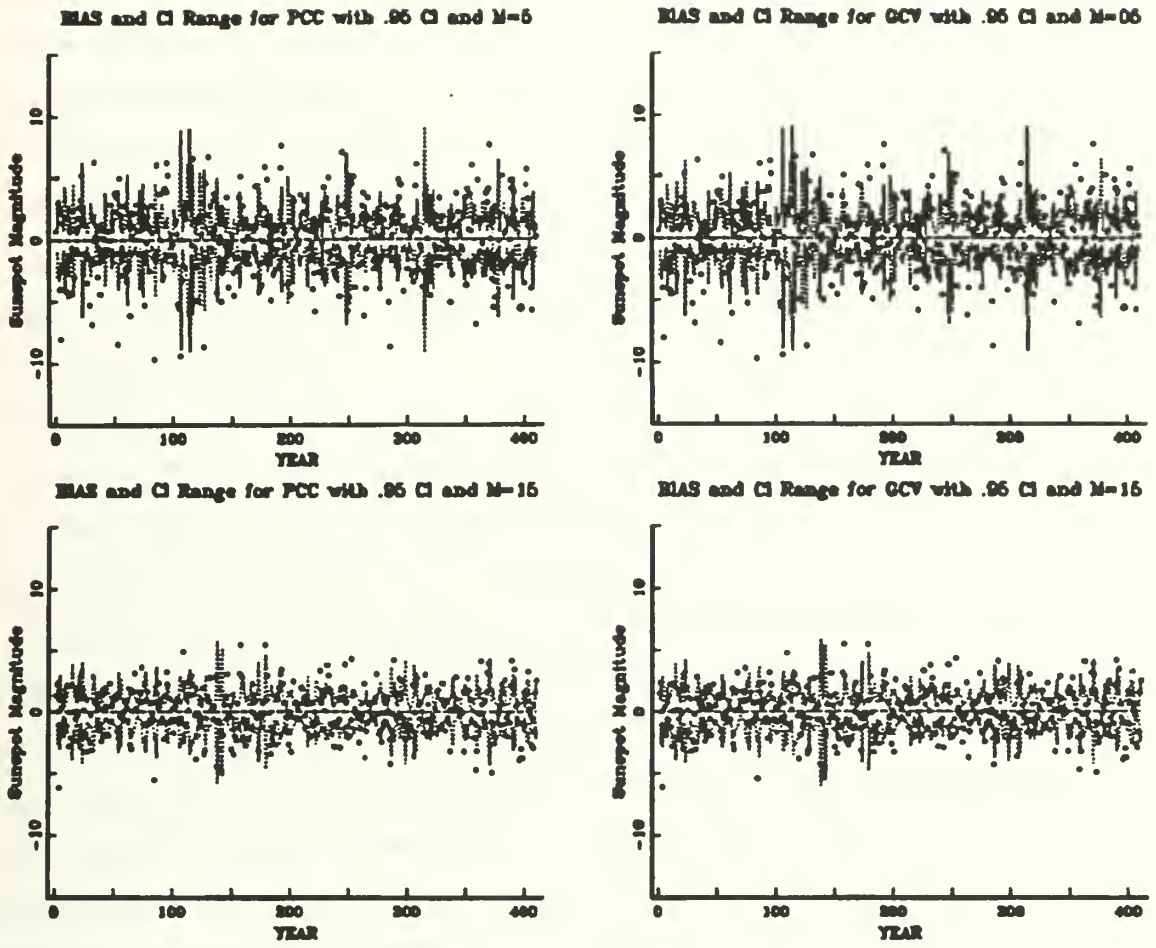

Figure 55. SIMULATION of the LIMIT CYCLE from ASTAR MODEL 9: The bias (points) and a 95% confidence interval centered about zero (lines) for the estimates of ASTAR Model 9's limit cycle. The simulation experiment used 50 simulations of 411 values of the limit cycle developed from ASTAR Model 9 of the Wolf Sunspot numbers with additive N(0,1) noise. The plots in this figure are for the *PC* [left] and GCV^* [right] model selection criteria using $M = 5$ [top] and $M = 15$ [bottom], the maximum number of subregions permitted in the forward step of the MARS algorithm.

E. SMASTAR MODELING OF THE VATNSDALSA RIVERFLOW USING MARS 3.0

In Chapter's II and III the GCV^* criterion in MARS 2.0 was used to develop models of the Sunspot numbers and the Vatnsdalsa riverflow system. Chapter IV discussed the MARS 3.0 program and its modifications to facilitate time series analysis. One of the modifications is an input variable, MSC, that allows the selection of different model selection criteria for use in the MARS algorithm. In this section the objective is three-fold; first to discuss the use of the model selection criteria investigated in Sections C and D of this chapter in conjunction with the new capabilities of MARS 3.0, second to determine if the model selection criterion in MARS 3.0 can be used to improve Model ICE486 of the Vatnsdalsa riverflow system, and finally to study the performance of the model selection criterion in a more difficult setting. Note that the $AIC2$ criterion was not investigated in this section due to its poor performance in Section D of this chapter.

Using MARS 3.0 and the GCV^*, AIC, PC and SC model selection criteria, 2 Year SMASTAR Models of the Vatnsdalsa Riverflow were developed using 731 days (2 years) of riverflow for model development and the remaining 355 days for prediction. Models using each criterion were permitted to form 1, 2, and 3-way interactions during a maximum of M = 25 and 50 forward steps of the MARS algorithm as compared to the 10 to 20 forward steps permitted during the Vatnsdalsa riverflow modeling discussed in Chapter III. The minimum span between threshold values for a single predictor variable was 50 and 75 values. The SMASTAR models were developed using a total of 27 predictor variables (9 lagged predictor variables for each time series); lagged riverflow $X_{\tau-1}$ to $X_{\tau-9}$, lagged precipitation $Y_{\tau-1}$ to $Y_{\tau-9}$, with and without the natural log transformation $Y^*_{\tau-i} = ln\,(1 + Y_{\tau-i})$, and lagged temperature $Z_{\tau-1}$ to $Z_{\tau-9}$. The first 9 data values of each time series were used for initialization.

Analysis of the fitted values and residuals of the models selected by the GCV^*, AIC and PC criteria indicate that these model selection criteria tend to create very large models for the riverflow if the number of forward steps of the MARS algorithm is set high i.e., when compared to the SC criterion, the GCV^*, AIC and PC criteria do not eliminate many terms from the SMASTAR model during the backward step of the MARS algorithm. One result of these large models is that the fitted residuals tend to have significant autocorrelation,

possibly due to overfitting. Also, when used for prediction, models developed using the GCV^*, PC and AIC criteria tend to have unpredictable results i.e., violent changes in behavior that at times can lead to negative riverflow (again this may be due to overfitting).

Use of the GCV^*, PC and AIC criteria necessitate judicious use of the model parameter that sets the maximum number of forward steps in the MARS algorithm. For example, Model ICE486 developed in Chapter III using the GCV^* criterion had 13 model terms during only 15 forward steps of the MARS algorithm. Yet in this experiment a model developed using GCV^* had 44 model terms during $M = 50$ forward steps of the MARS algorithm. When compared to the models developed using the SC criterion, the GCV^* criterion appears to over parameterize the SMASTAR model.

This experiment also indicates that the final size of the model is due, in part, to the relationship between the two parts of each model selection criterion (model complexity and model lack-of-fit). Recall that the apparent model over-parameterization of linear time series models by the AIC criterion led to the development of the SC criterion (Schwarz, 1978), which increases the weight of the model complexity function by a factor of .5 $ln(N)$. Thus for a given value of N, to add a term to a SMASTAR model and improve (decrease) the SC criterion's 'score' requires a greater decrease in the model's lack-of-fit than required using the AIC criterion.

Equation (67) details SMASTAR Model ICE SC160 for the Vatnsdalsa riverflow for the years 1972 and 1973, developed using the SC model selection criterion. Model ICE SC160 for the Vatnsdalsa riverflow uses the natural log transformed precipitation and was permitted to form 1, 2, and 3-way interactions during a maximum of $M = 50$ forward steps of the forward step MARS algorithm. The minimum span between threshold values for a single predictor variable was 50 data values. SMASTAR Model ICE SC160, which should be compared to SMASTAR Model ICE486 (equation 54), is

$$
\widehat{X}_\tau = \left\{
\begin{array}{ll}
4.13 & +0.1940(X_{\tau-1}-3.98)_+(X_{\tau-2}-8.36)_+ \\
 & -0.0003(X_{\tau-1}-3.98)_+(X_{\tau-2}-8.36)_+(X_{\tau-6}-3.98)_+ \\
 & \left\{
\begin{array}{l}
-4.572(.182-Y^*_{\tau-2})_+ \\
+0.574Y^*_{\tau-1}(Y^*_{\tau-4}-1.53)_+ \\
-0.662(X_{\tau-3}-9.02)_+(1.53-Y^*_{\tau-4})_+ \\
\left\{
\begin{array}{l}
+1.211(X_{\tau-1}-3.98)_+(Y^*_{\tau-2}-.262)_+ \\
-1.321(X_{\tau-1}-3.98)_+(Y^*_{\tau-2}-1.07)_+ \\
+1.035(X_{\tau-1}-3.98)_+(1.07-Y^*_{\tau-2})_+ \\
-0.145(X_{\tau-1}-3.98)_+(1.07-Y^*_{\tau-2})_+(Y^*_{\tau-4}-1.03)_+
\end{array}
\right. \\
\left\{
\begin{array}{l}
+0.085(X_{\tau-1}-3.98)_+(8.36-X_{\tau-2})_+Y^*_{\tau-1} \\
+0.023(X_{\tau-1}-3.98)_+(X_{\tau-2}-8.36)_+(.262-Y^*_{\tau-8})_+
\end{array}
\right.
\end{array}
\right. \\
 & \left\{
\begin{array}{l}
\left\{
\begin{array}{l}
+.0146(X_{\tau-1}-3.98)_+(3.00-Z_{\tau-1})_+ \\
-.0035(X_{\tau-1}-3.98)_+(X_{\tau-2}-3.98)_+(3.00-Z_{\tau-1})_+ \\
-.0176(X_{\tau-1}-3.98)_+(Z_{\tau-1}-3.00)_+(4.80-Z_{\tau-7})_+
\end{array}
\right. \\
\left\{
\begin{array}{l}
+.0084(X_{\tau-1}-3.98)_+(Z_{\tau-1}+1.60)_+(3.50-Z_{\tau-6})_+ \\
-.0033(X_{\tau-1}-3.98)_+(X_{\tau-3}-3.98)_+(Z_{\tau-1}+1.60)_+
\end{array}
\right. \\
-.0081(X_{\tau-1}-3.98)_+(X_{\tau-2}-8.36)_+(Z_{\tau-2}+22.4)_+
\end{array}
\right.
\end{array}
\right. \qquad (67)
$$

Model ICE SC160 (Figure 56) has 32 parameters that includes 19 terms (a model constant term and 1 one-way, 7 two-way and 10 three-way interactions) and 13 threshold values (1 each on $X_{\tau-2}$, $X_{\tau-3}$, $X_{\tau-4}$, $Y^*_{\tau-8}$, $Z_{\tau-6}$, $Z_{\tau-7}$), 2 on $Y^*_{\tau-4}$ and $Z_{\tau-1}$ and 3 on $Y^*_{\tau-2}$. The standard error of the fitted residuals is $\sigma_\epsilon = 1.10m^3/sec.$ for Model ICE SC160 versus $\sigma_\epsilon = 1.27m^3/sec.$ for Model ICE496 developed using GCV^*. Figure 56 shows plots of the fitted values and residuals of Model ICE SC160 for the Vatnsdalsa riverflow data during 1972 and 1973.

Model ICE SC160, Figure 56, appears to equally overfit and underfit the peaks and troughs as it captures the general structure of the riverflow data. Analysis of the normal probability plot (not shown) shows that the fitted residuals are still slightly skewed with

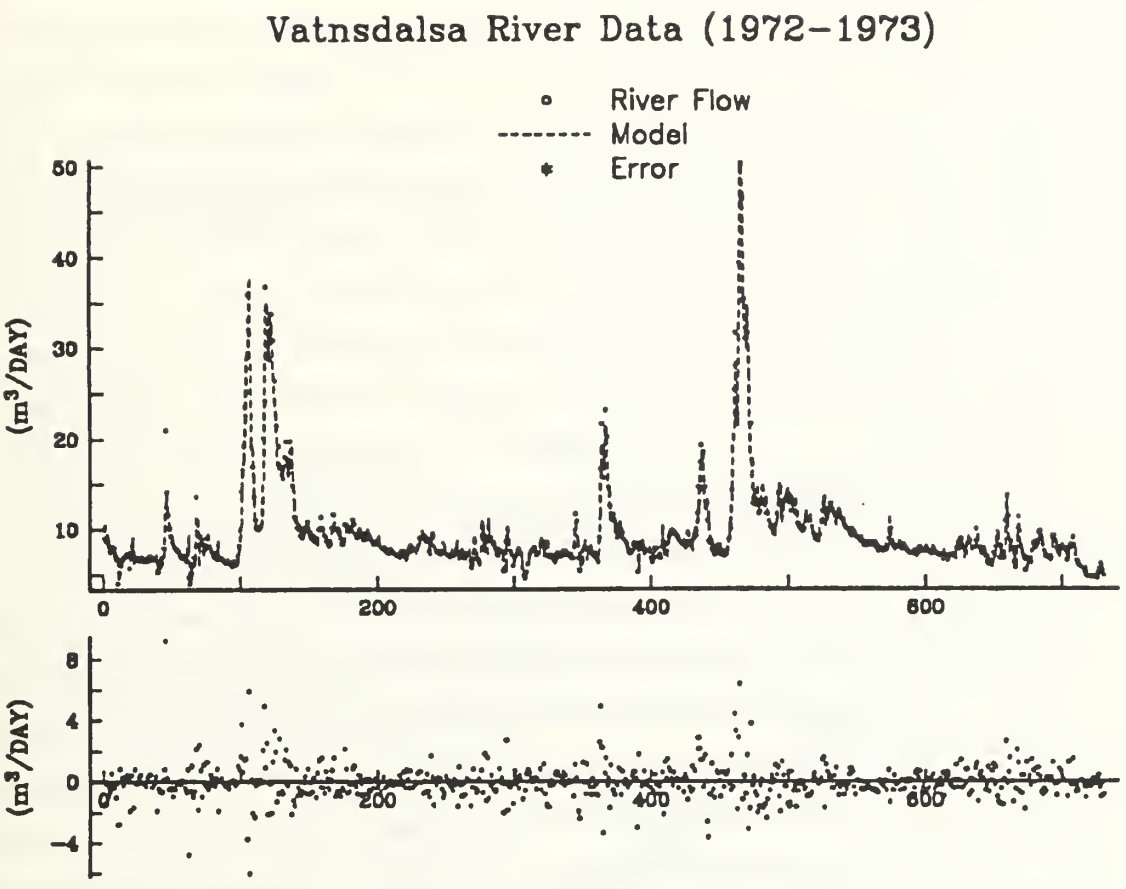

Figure 56. The Vatnsdalsa riverflow data for years 1972 and 1973 versus the fitted values (top) and residuals (bottom) for SMASTAR Model ICE SC160. The SMASTAR model for the riverflow at time τ, X_τ, was a function of lagged riverflow $X_{\tau-1}$ to $X_{\tau-9}$, lagged precipitation $Y^*_{\tau-1}$ to $Y^*_{\tau-9}$, i.e., the natural log transformation $Y^*_{\tau-i} = ln(1+Y_{\tau-i})$, and lagged temperature $Z_{\tau-1}$ to $Z_{\tau-9}$. Model ICE SC160 has 32 parameters that includes 19 terms (a model constant term and 1 one-way, 7 two-way and 10 three-way interactions) and 13 threshold values (1 each on $X_{\tau-2}$, $X_{\tau-3}$, $X_{\tau-4}$, $Y^*_{\tau-8}$, $Z_{\tau-6}$, $Z_{\tau-7}$), 2 on $Y^*_{\tau-4}$ and $Z_{\tau-1}$ and 3 on $Y^*_{\tau-2}$. The standard error of the fitted residuals is $\sigma_\epsilon = 1.10 m^3/sec$. The initial nine values of each time series were used to initialize the model.

the extremely heavy tails that have occurred with previous modeling efforts of this type riverflow data. The fitted residual autocorrelation function and estimated normalized periodogram plots are shown at Figure 57. As with Model ICE486 using GCV^* (Figure 31) the autocorrelation function for the fitted residuals reveals no evidence of short term autocorrelation. Also, as with Model ICE486, we could consider the residuals independent if they were normally distributed because the normalized cumulative spectrum of the fitted residuals falls entirely within the 90% K-S bounds from the cumulative spectrum for white noise. However, again the fitted residuals display a pattern of high residual values during periods of high riverflow (Figure 56), evidence of the non-normality of the fitted residuals.

To investigate the predictive performance of Model ICE SC160, developed and discussed above, Model ICE SC160 and the riverflow, precipitation and temperature data during the year 1974 was used to perform a 1 day forward-step ahead prediction of the Vatnsdalsa riverflow. Overall the predictions of Model ICE SC160 are only slightly different than the 1 day forward-step ahead predictions of Model ICE486 using GCV^*. Figures 58-59 contain plots of the actual versus 1 day forward-step ahead predictions of Model ICE SC160 and the fitted residuals for the Vatnsdalsa riverflow during the year 1974. Again, the 1 day forward-step ahead predictions were performed using coefficient updating and a fixed coefficient model. In both cases the model predictions react very well to both the extreme spring transition and low riverflow that occurs later in the year. However, as expected the 1 day forward-step ahead predictions using coefficient updating (Figure 58) are an improvement over the 1 day forward-step ahead predictions using the fixed coefficient model (Figures 59). The standard error of the fitted residuals using coefficient updating is $\sigma_\epsilon = 2.08\ m^3/sec.$ for Model ICE SC160 using SC versus $\sigma_\epsilon = 2.11\ m^3/sec.$ for Model ICE486 using GCV^*. The standard error of the fitted residuals using the fixed coefficient model is $\sigma_\epsilon = 2.67\ m^3/sec.$ for Model ICE SC160 using SC versus $\sigma_\epsilon = 2.37\ m^3/sec.$ for Model ICE486 using GCV^*. The predictive capability of the two models is similar. Note that Model ICE SC160 has a slightly smaller fitted residual variance than Model ICE486 for the coefficient updating method while the opposite is true for the fixed coefficient method. However, Model ICE486 was developed in a restrictive environment with only $M = 15$ forward steps of the MARS algorithm while Model ICE SC160 was developed in an unrestrictive and thus preferable environment with $M = 50$ forward steps of the MARS algorithm.

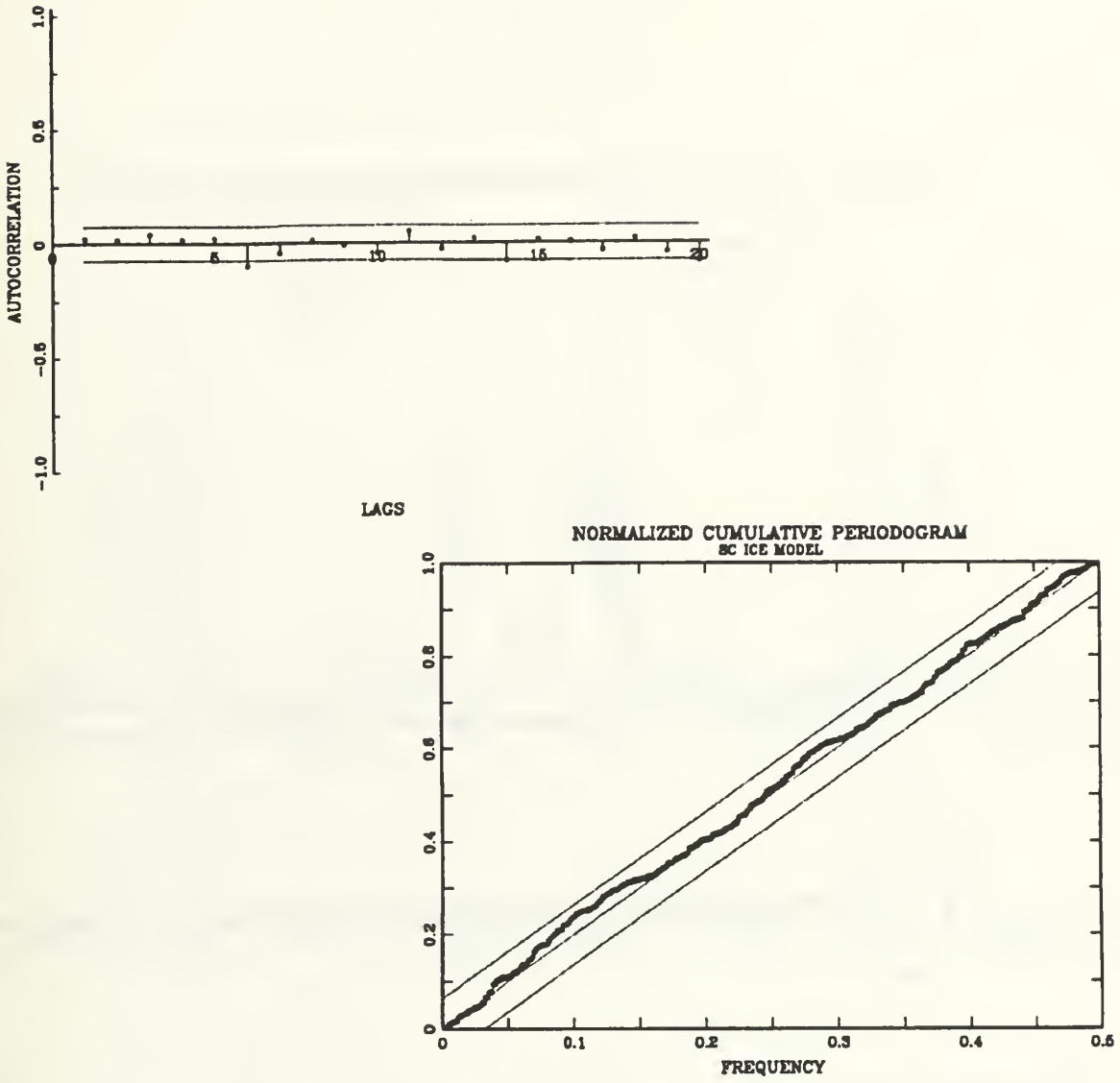

Figure 57. Fitted Residual Plots from SMASTAR Model ICE SC160. The autocorrelation function (first 20 lags) [top] and the normalized cumulative periodogram [bottom] of the fitted residuals from SMASTAR Model ICE SC160 of the Vatnsdalsa River system for the period 1972-1973. The top plot, with approximate 95% individual confidence bounds, shows that no apparent autocorrelation exists in the fitted residuals. Also the K-S bounds in the normalized cumulative periodogram plot indicates no departure from a flat spectrum, so that if the residuals are normally distributed, the residuals are independent.

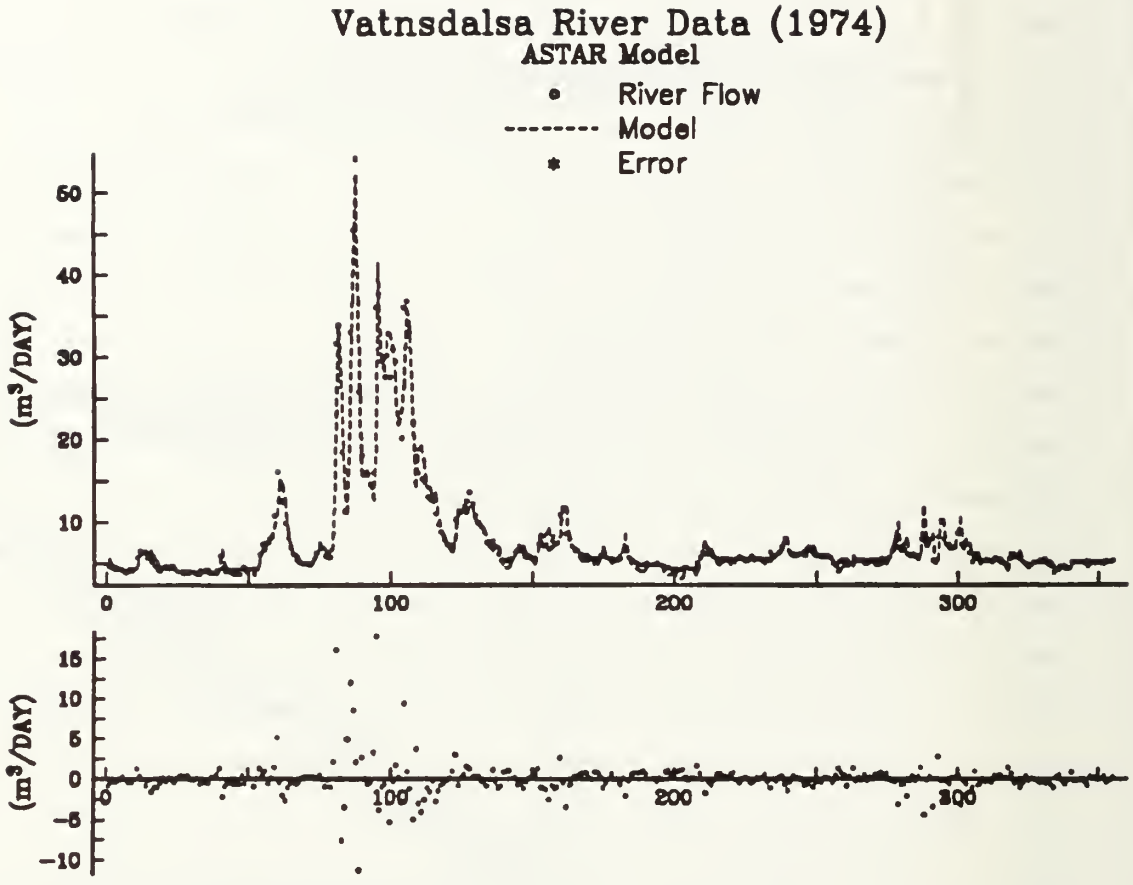

Figure 58. The actual versus 1-step ahead predictions and errors from MODEL ICE SC160 for the Vatnsdalsa riverflow data (1974) with coefficient updating (coefficient update). The standard error of the fitted residuals $\sigma_\epsilon = 2.08\ m^3/sec.$ for Model ICE SC160 versus $\sigma_\epsilon = 2.11\ m^3/sec.$ for Model ICE486 using GCV^*.

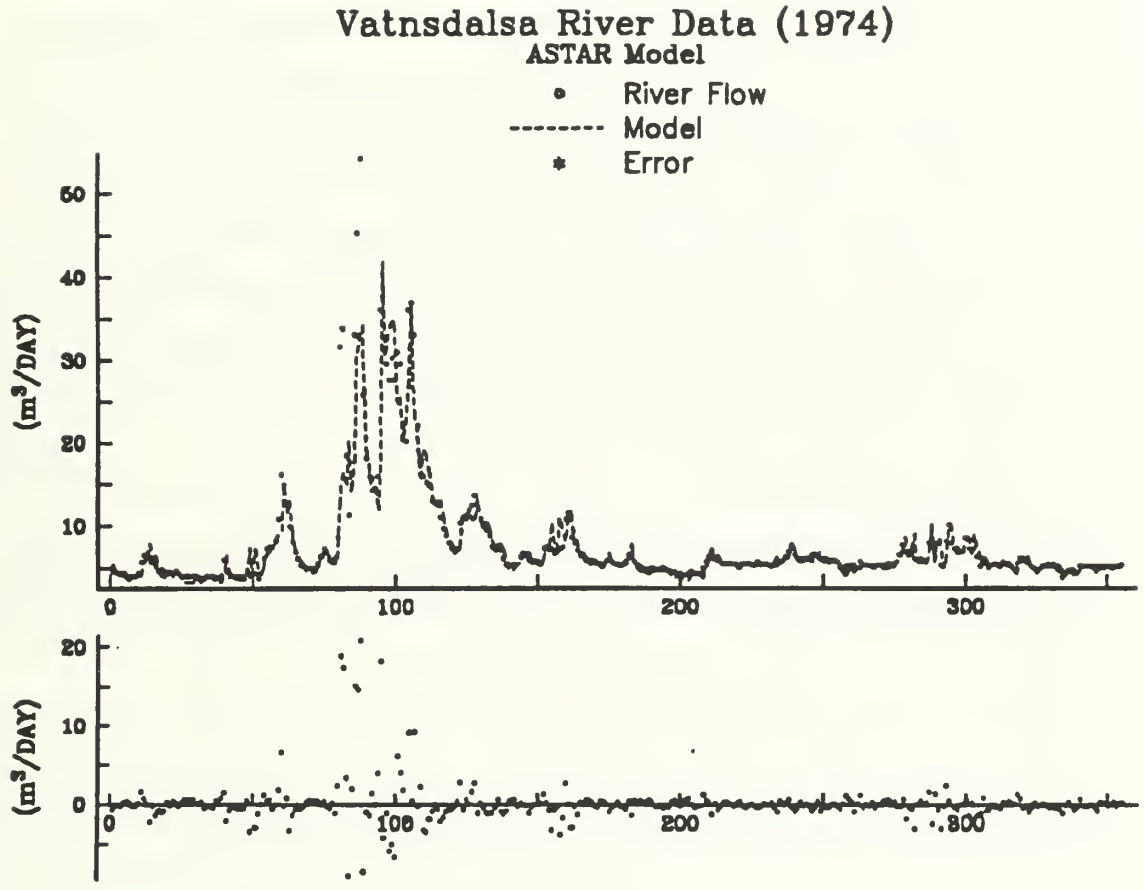

Figure 59. The actual versus 1-step ahead predictions and errors from MODEL ICE 160 for the Vatnsdalsa riverflow data (1974) without coefficient updating (fixed model). The standard error of the fitted residuals $\sigma_\epsilon = 2.67\ m^3/sec.$ for Model ICE SC160 versus $\sigma_\epsilon = 2.36\ m^3/sec.$ for Model ICE486 using GCV^*.

F. SUMMARY

This chapter examined the problem of model dimension and variable selection when using adaptive regression splines to develop a nonlinear autoregressive model for a univariate or semi-multivariate time series system. Five model selection criteria, GCV^*, AIC, $AIC2$, PC and SC, were examined to determine which performed best within MARS. The results indicate that SC is the best model selection criterion for use in MARS for a time series setting. The SC criterion (Schwarz, 1978; Rissanen, 1978) consistently performed well for all experiments conducted in this chapter and appears to best accommodate the forward and backward stepwise MARS strategy for model development. In contrast, the AIC criterion appeared to over parameterize models which agrees with findings by Schwarz (1978) and others. Model over-parameterization was also a characteristic of the PC and GCV^* criteria when modeling the Vatnsdalsa riverflow in an 'unrestricted' environment (large M, the number of forward steps in the MARS algorithm). The $AIC2$ criterion performed very poorly during simulations of the fitted values and limit cycle of ASTAR Model 9 of the Wolf sunspot numbers.

Thus the SC criterion is recommended for model selection when MARS is applied in a time series setting.

VI. THESIS SUMMARY

MARS is a new nonparametric regression modeling methodology, due to Friedman, that utilizes low-order regression spline modeling and a modified recursive partitioning strategy to exploit the localized low-dimensional behavior of the data used to construct $\hat{f}(\boldsymbol{x})$. Given a set of predictor variables, MARS fits a model in the form of an expansion in product spline basis functions of predictors chosen during a forward and backward recursive partitioning strategy. Although MARS is a computationally intensive regression modeling methodology, it provides a systematic (automatic) method for deriving nonlinear threshold models for high-dimensional data. The MARS models are naturally continuous in the domain of the predictor variables, and can have multiple partitions and predictor variable interactions.

Within MARS by letting the predictor variables for the τth value in a time series $\{X_\tau\}$ be its lagged values, i.e., $X_{\tau-1}$, $X_{\tau-2}$, ..., $X_{\tau-p}$, one obtains an adaptive spline threshold autoregressive (ASTAR) model, a new method for systematic nonlinear modeling of time series that extends the threshold autoregressive (TAR) model due to Tong (1985). Simulations of autoregressive and nonlinear threshold models are used to show the ability of ASTAR to model simple time series. A significant feature of ASTAR models when modeling time series data with periodic behavior is its ability to produce continuous models with underlying sustained oscillations (limit cycles). The initial analysis of the yearly Wolf sunspot numbers (1700-1890) and (1700-1920) using ASTAR produced several models with underlying limit cycles. When used to predict the yearly sunspot numbers (1921-1955), the ASTAR models are a significant improvement over existing Threshold and Bilinear models.

Within MARS by letting the predictor variables be not only the lagged values of the time series being modeled and predicted, but also the lagged values of other related time series, results in a semi-multivariate adaptive spline threshold autoregressive (SMASTAR) model. This investigation indicates that SMASTAR models appear well suited for taking into account the complex nonlinear interactions among multivariate, cross-correlated, lagged predictor variables of a time series system. Using the Vatnsdalsa riverflow system

as an example, Tong et al. (1985) showed that normal linear autoregressive models were incapable of modeling the complexities for this type time series system. Also the methodology and structure of semi-multivariate TAR models appear incapable of capturing these complexities in a parsimonious model. However, the SMASTAR model appears to consider the complex relationships between the cross-correlated predictor variables, and seems capable of providing semi-multivariate nonlinear time series models for prediction, even in non-normal situations such as riverflow data. SMASTAR model development, although computationally intensive, is also quite systematic.

An important aspect of any overall regression modeling effort is the interpretation and analysis of a regression model. However, the functional form of an ASTAR model, with its combination of different predictor variables and multiple threshold values, makes its straightforward interpretation and analysis difficult. In this regard a graphical representation was developed to permit the interpretation and analysis of ASTAR models. It was shown that this graphical representation can be used to analyze the use for and contribution of each of the terms in an ASTAR model. The extension of this graphical representation to SMASTAR models is difficult. However, it was shown that the tree-like structure of a MARS model can be used to analyze the use for and contribution of each of the terms in a SMASTAR model.

The current model selection criterion in MARS is GCV^*, a modified form of generalized cross validation. However, other model selection criterion, such as Akaike's Information Criterion (AIC), have been suggested for model selection in the standard linear time series setting. In this regard, simulations were used to investigate GCV^* and several other model selection criterion for use within MARS. The results indicate that the Schwarz-Rissanen's SC criterion and Amemiya's PC criterion improve the model selection over GCV^* when MARS is used in a time series setting. The simulation experiments identified the potential for over parameterization by AIC that has been identified by Schwarz and others. The PC and GCV^* criteria also appeared to create unnecessarily large models (lack of parsimony) when used for model selection of the Vatnsdalsa riverflow system. *Thus in a time series setting the SC criterion is the recommended model selection criterion for use within MARS.*

The application of MARS for nonlinear modeling of univariate and semi-multivariate time series systems is a new and exciting methodology. However, there is still a need for

additional investigation of many aspects of this application. Additional comments, questions and areas for further research include;

1. It is important to note that as with any regression or time series modeling effort, one can never be sure that one has all the relevant predictor variables. However, MARS version 3.0 has been modified and Fortran programs written to permit analysis of univariate and multivariate time series systems. The modifications include the ability to select from one of several model selection criterion that have been proposed for use in a time series setting. What other structural and methodological modifications are needed to improve nonlinear modeling of time series using MARS?

2. As discussed in Chapter IV the MARS 3.0 program permits the use of categorical variables such as wind velocity or circular wind direction. The use of categorical variables in SMASTAR time series models has not been investigated. However, this time series modification to MARS appears to provide an innovative approach for including influential lagged categorical variables.

3. A constant source of concern when modeling time series data is that of variance homogeneity and independence of the error term in the model

$$X_\tau = f(X_{\tau-1}, \ldots, X_{\tau-p}) + \varepsilon_\tau.$$

   Residual analysis of the yearly sunspot data model supported this assumption of homeoscedascity and independence. If not, an initial attempt to overcome this would be, for positive data, to use a log transformation as was done for the precipitation data from the Vatnsdalsa riverflow system. However, there is no guarantee that this would work; for instance if the data were generated by an ARCH model (Tong, 1990 pp. 116-117) then a simple transformation of the data, such as a log transformation, would not work. Other questions involve normality of the errors.

4. MARS selects a model using exhaustive search and is a computer intensive methodology. Note that it is quite systematic and interactions and thresholds are selected by the stepwise methodology. Also some of the predictor variables used as candidates for the model may be rejected during the forward step of the MARS algorithm, i.e, MARS uses subset selection of the available predictor variables. Tsay (1989) has developed procedures for threshold variable selection and a statistic to test threshold values within the framework of TAR. These statistics, and graphical methods that parallel the methods discussed for the Wolf sunspot numbers in Chapter II need to be incorporated, in a fairly automatic way, in the MARS methodology for time series analysis.

5. Can MARS be used to measure the degree of nonlinearity of a time series system?

6. An issue of concern in the general application of MARS is the method for determining the 'correct' number of degrees-of-freedom to charge for variable and threshold value selection. This issue is even less clear across the lagged predictor variables of a univariate or semi-multivariate time series system.

7. As discussed in Chapter IV there are several model parameters that must be set to initialize the MARS algorithm. It is preferable to set the parameters to be as unrestrictive as possible and permit the model selection criterion and the data to determine the final model form. In this regard, the guidance (obtained mostly by practical experience) offered by Friedman (1991) for setting the MARS model parameters in the general setting appears very robust. However, the practical application of MARS to time series has been limited to this investigation.

8. The use of ASTAR and SMASTAR models for modeling and analysis followed by the simulation of complex, nonlinear systems is discussed in Lewis and Stevens (1990). Frequently, individual inputs of complex global system models are the result of the analysis of subcomponent systems. If the subcomponent systems are time series systems then the reduction of these nonlinear time series systems to a tractable model form such as provided by ASTAR and SMASTAR time series models may enhance the efficiency and accuracy of global system inputs.

APPENDIX A. FORTRAN BATCH FILE FOR DEVELOPING ASTAR AND SMASTAR TIME SERIES MODELS USING THE MARS 3.0 PROGRAM

```
@ECHO OFF
This file is marstsa.bat.                              01 July 1991
*****************************************************************
PC BATCH FILE FOR FOR INITIATING MARS3.0a TIME SERIES RUNS FOR
MICROWAY NDP FORTRAN 2.1.4 UNDER DOS USING THE WEITEK
COPROCESSOR.  CHANGE THE -N4 PARAMETER TO -N2 IN THE MARCL.BAT
FILE AND RECOMPILE THE MARS3.0A, MARSBLDa AND MARSDRVa FORTRAN
PROGRAMS TO USE THE INTEL OR CYRIX COPROCESSORS.  Does not run
under Ver 2.0.6 of MICROWAY NDP FORTRAN, and has not been tried
under their later versions.  At least Version 3 exists, maybe
Version 4.
*****************************************************************
J. STEVENS - L. URIBE - P.A.W. LEWIS.   e-mail 1526P@NAVPGS.BITNET
*****************************************************************

THIS EXEC PREPARES THE INPUT REGRESSION MATRIX FOR FRIEDMAN'S
MARS3.0a OUT OF 1, 2 OR 3 TIME SERIES USING THE MARSBLDa FORTRAN
PROGRAM.  AFTER THAT IT CALLS THE MARSDRVa FORTRAN PROGRAM WHICH
PERFORMS THE MARS REGRESSION, FIRST COMPUTING ALL THE ARRAY
SPACE ALLOCATIONS NEEDED IN MARS3.0a IN AN AUTOMATED WAY.  THIS
RELIEVES THE USER FROM THE BURDEN OF SUCH GUESSWORK.  THE ONLY
VALUE THAT MAY NEED ADJUSTMENT FROM THE USER IS ON THE 2ND LINE
OF MARSDRVa FORTRAN PROGRAM, WHERE THE SIZE FOR THE PARAMETER NV
APPEARS.  THIS PARAMETER IS USED TO INCREASE OR REDUCE THE
AMOUNT OF MEMORY AVAILABLE FOR MARS3.0a.  IN SUCH CASE MARSDRVa
NEEDS TO BE RECOMPILED PRIOR TO RUNNING THIS BATCH EXEC.

PROGRAM MARSBLDa PROMPTS FOR THE NAMES OF UP TO 3 TIME SERIES FILE
NAMES.  THE 1ST ONE HAS THE TIME SERIES BEING PREDICTED, FROM LAGGED
VALUES OF ITSELF AND it must always be PRESENT.  THE OTHER 2 TIME
SERIES ARE OPTIONAL PREDICTOR TIME SERIES. PRESS THE
<ENTER> KEY ALONE WHEN  ANY OF THESE 2 TIME SERIES IS NOT USED.

OUTPUT RESULTS FROM MARS3.0a AND AN ADDITIONAL INTERACTIONS REPORT
```

```
APPEARS ON FILE UNIT 06 (MARS3.OUT)

INPUT FILE(S) FORMAT:  (See Definitions Below)
RECORD  1:     N,P,MI,NK,NGC,NGS,M,ICX,MS,DF,MSC
RECORD  2:     LX(I),  I=1,P
RECORD  3:     LAG(I), I=1,P
RECORDS 4-END: TIMES SERIES VALUES

ALL THE ABOVE INFORMATION IS ENTERED IN FREE FORMAT (JUST ONE OR
MORE SPACES BETWEEN VALUES).  RECORD 2 AND RECORD 3 CAN BE MULTIPLE
RECORDS THEMSELVES WHEN P IS LARGE.  THE ARRAYS CAN BE ENTERED
FOR EXMPLE 20 VALUES PER LINE.

PARAMETER DEFINITIONS:

      N=NO. OF VALUES IN THE TIME SERIES (ALL 3 MUST BE EQUAL)
      P=NO. OF PREDICTORS FROM THIS TIME SERIES
      MI=MAX. NO OF INTERACTIONS
      NK=MAX. NO. OF BASIS POINTS
      NGC=NO. OF RASTER POINTS FOR PLOTTING (SET TO 0 FOR NO PLOT)
      NGS=NO. OF R.P. ON EACH AXIS FOR PLOTS(          "          )
      M=MODEL FLAG: 1=PLOT PIECEWISE LINEAR, 2=PLOT PIECEWISE CUBIC
      ICX=CONVEX HULL FLAG:  0=PLOT SURFACE OVER ENTIRE RANGE OF ARGS.
                            >0=PLOT SURF. OVER INSIDE CONVEX HULL
      MS=MIN. SPAN (MIN NO. OBSERVATIONS BETWEEN KNOTS)
      DF=NO. OF DEGREES OF FREEDOM
      MSC=MODEL SELECTION CRITERIA (1=GCV, 2=AIC, 3=PC, 4=SC)

ENTER 0 FOR THOSE PARAMETER VALUES NOT APPLICABLE TO A GIVEN RUN.

      LX=PREDICTOR VAR. FLAG: 0=EXCLUDE VARIABLE FROM MODEL
                              1=ORDERABLE VARIABLE. NO RESTRICTION
                              2=ORDERABLE VAR. ADDITIVE. NO INTERACTS.
                              3=ORDERABLE VARIABLE LINEAR ONLY.
                             -1=CATEGORICAL VAR. NO RESTRICTION.
                             -2=CATEGORICAL VAR. ADDITIVE. NO INTERACTS
      LAG=LAGS TO USE TO GENERATE PREDICTORS FROM THIS TIME SERIES.

      TIME SERIES VALUES: THEY FOLLOW IN FREE FORMAT, AS MANY PER
                          RECORD AS DESIRED.

THE TOTAL NO. OF PREDICTORS IS THE SUM OF THE PREDICTORS FOR EACH
INPUT TIME SERIES.
```

```
NOTE THAT ALL VALUES ON RECORD 1 OF THE 3 FILES MUST BE THE SAME
EXCEPT FOR P THE NUMBER OF PREDICTORS.

BREAK ON
@ECHO On
SET SAVPTH=%PATH%
Save old Path, and below create one for the MARS run.
PATH c:\;c:\dos401;D:\;d:\NDP20;d:\NDP20\MARSNEW
RUN386 MARSBLDa
RUN386 MARSDRVa
PATH=%SAVPTH%
```

APPENDIX B. NDP FORTRAN PROGRAM FOR BUILDING THE INPUT TO THE MARS 3.0 PROGRAM FOR ASTAR AND SMASTAR TIME SERIES MODEL DEVELOPMENT

```
c     This is the MARSBLD.F Fortran Program              01 July 1991
c   ****************************************************************
c   J. STEVENS - L. URIBE - P.A.W. LEWIS.   e-mail 1526P@NAVPGS.BITNET
c   ****************************************************************
C --- BUILD THE STD INPUT DATA FILE FOR MARSDRV WITH 1 TO 3 SERIES
C --- CALLS FOR UP TO 3 INPUT FILES, WHOSE FORM IS GIVEN IN THE
C --- MARSTSA.BAT FILE WHICH CALLS THIS FILE.
C --- PARAMETER DEFINITIONS GIVEN IN THE MARSTSA.BAT FILE
      INTEGER P, P1,P2,P3
      PARAMETER(MXP=100,MXN=10000)
      INTEGER LX1(MXP),LX2(MXP),LX3(MXP), LG1(MXP),LG2(MXP),LG3(MXP)
      REAL X(MXN,MXP),Y(MXN),W(MXN),X1(MXN),X2(MXN),X3(MXN)
      CHARACTER*12 FN1,FN2,FN3, FOUT
      DATA P1,P2,P3, NXX,NX2,NX3 /0,0,0,0,0,0/
C
*  DATA INPUT LINE 1 PARAMETERS --  NXX,P,MI,NK,NGC,NGS,M,ICX,MS,DF,MSC
*             LINE 2            --  LX
*             LINE 3            --  DESIRED LAG VARIABLES. ORDERED
*             REST OF FILE      --  TIME SERIES
C

      FOUT='MARS30a.DAT'
      OPEN(10,FILE=FOUT)
C
      WRITE(6,*) 'UNDER IBM CMS FILE NAMES MUST BEGIN WITH A / '
      WRITE(6,*) 'ENTER 1ST TIME SERIES FILE NAME (IN QUOTES):'
      READ(5,*) FN1
      OPEN(7,FILE=FN1,ERR=999)
      CALL GETDATA(7, NXX, P1,MI,NK,NGC,NGS,M,ICX,MS,DF,LX1,LG1,X1,
     *      MXN,MXP,MSC)
C
      WRITE(6,*) 'ENTER 2ND TIME SERIES FILE NAME (IN QUOTES):'
      WRITE(6,*) 'IF NOT APPLICABLE JUST TYPE ONE SPACE IN QUOTES'
      READ(5,*) FN2
      IF(FN2.NE.'/' .AND. FN2.NE.' ') THEN
        OPEN(8,FILE=FN2,ERR=999)
```

```
        CALL GETDATA(8, NX2, P2, I, I,  I,  I,I,  I, I, A,LX2,LG2,X2,
     *      MXN,MXP,I)
       ENDIF
       WRITE(6,*) 'ENTER 3RD TIME SERIES FILE NAME (IN QUOTES):'
       WRITE(6,*) 'IF NOT APPLICABLE JUST TYPE ONE SPACE IN QUOTES'
       READ(5,*) FN3
       IF(FN3.NE.'/' .AND. FN3.NE.' ') THEN
         OPEN(9,FILE=FN3,ERR=999)
         CALL GETDATA(9, NX3, P3, I, I,  I,  I,I,  I, I, A,LX3,LG3,X3,
     *      MXN,MXP,I)
       ENDIF
       if(nxx.eq.0 .or.   nxx.ne.nx2 .and.nx2.gt.0  .or.
     *                    nxx.ne.nx3 .and.nx3.gt.0) then
           write(6,*) 'series are not of the same length',nxx,nx2,nx3
           close(10)
           stop
       endif
C
       l2=0
       l3=0
       if(p2.gt.0) l2=lg2(p2)
       if(p3.gt.0) l3=lg3(p3)
       LP=MAX(LG1(P1), L2, L3)
       N=NXX-LP
       P=P1+P2+P3
C --   WEIGHTS W.  RESPONSE Y BUILT FROM 1ST TIME SERIES
       DO 100 II=1,N
        W(II)=1.
        Y(II)=X1(II+LP)
100    CONTINUE
C
C --- BUILD THE REGRESSION X MATRIX
       DO 101 II=1,N
         DO 102 JJ=1,P1
           X(II,JJ)=X1(II+LP-LG1(JJ))
102      CONTINUE
101    CONTINUE
C
       DO 103 II=1,N
         DO 104 JJ=1,P2
           X(II,JJ+P1)=X2(II+LP-LG2(JJ))
104      CONTINUE
103    CONTINUE
```

```
C
      DO 105 II=1,N
        DO 106 JJ=1,P3
          X(II,JJ+P1+P2)=X3(II+LP-LG3(JJ))
106     CONTINUE
105   CONTINUE
C
C --- BUILD MARSDRV INPUT FILE
      WRITE(10,114) N,P1,P2,P3,MI,NK,NGC,NGS,M,ICX,MS,DF,MSC
      WRITE(10,111) (LX1(I),I=1,P1), (LX2(I),I=1,P2), (LX3(I),I=1,P3)
      WRITE(10,116) (LG1(I),I=1,P1), (LG2(I),I=1,P2), (LG3(I),I=1,P3)
      WRITE(10,112) (W(I),I=1,N)
      DO 110 I= 1,N
       WRITE(10,112) (X(I,J),J=1,P), Y(I)
110   CONTINUE
      RETURN
999   CONTINUE
      WRITE(6,*) 'FILE NOT FOUND FOR THIS TIME SERIES'
      STOP
C
111   FORMAT(20I3)
112   FORMAT(14F10.5)
114   FORMAT(11I5,F5.1,I5)
116   FORMAT(20I5)
      END
C
      SUBROUTINE GETDATA(IU,NXX,P,MI,NK,NGC,NGS,M,ICX,MS,DF,LX,LAG,X,
     *                MXN,MXP,MSC)
      INTEGER LX(MXP),LAG(MXP), IU,P
      REAL X(MXN)
      READ(IU,*,END=100) NXX,P,MI,NK,NGC,NGS,M,ICX,MS,DF,MSC
      READ(IU,*,END=888) (LX(J),J=1,P)
      READ(IU,*,END=888) (LAG(J),J=1,P)
      READ(IU,*,END=888) (X(I), I=1,NXX)
      CLOSE(IU)
      DO 10 I=2,P
       IF(LAG(I).LE.LAG(I-1)) THEN
        WRITE(6,*) 'LAGS NOT IN ASCENDING ORDER OR DUPLICATE,UNIT=',IU
        STOP
       ENDIF
   10 CONTINUE
      return
  100 continue
```

```
c --- empty file
      nxx=0
      p=0
      RETURN
888   CONTINUE
      WRITE(6,*) 'FILE INCOMPLETE FOR THIS TIME SERIES'
      STOP
      END
```

APPENDIX C. NDP FORTRAN PROGRAM FOR EXECUTING THE MARS 3.0 PROGRAM

```
C This is the MARSDRVA.F FORTRAN file                    01 JULY 1991
C ***********************************************************************
C   J. STEVENS - L. URIBE - P.A.W. LEWIS.   e-mail 1526P@NAVPGS.BITNET
C ***********************************************************************
C DRIVER PROGRAM FOR RUNNING MARS 3.0. IT FIRST COMPUTES ALL THE ARRAY
c SPACE ALLOCATIONS, AND THEN RUNS THE MARS REGRESSION, USING THE
C INPUT PREPARED BY MARSBLDA.F WHICH IS CALLED BY MARSTSA.BAT.
      PARAMETER (NV=  20000)
C     SET UP WORKING STORAGE:
      REAL V(NV)
      INTEGER INTV(NV)
      CHARACTER*8 FIN
      EQUIVALENCE (V, INTV)
C
       OPEN(10,FILE='mars30a.DAT',ERR=999)
       OPEN(6,FILE='mars30a.out',ERR=999)
c lu  FIN='MARS30A'
c lu   OPEN(10,FILE=FIN // '.DAT',ERR=999)
c lu   OPEN(6, FILE=FIN // '.OUT')
      WRITE(6,'(/,''  DRIVER FOR MARS 3.0. '')')
C
C READ IN DATA:
C
      READ(10,*,END=999) N,NP1,NP2,NP3,MI,NK,NGC,NGS,M,ICX,MS,DF,MSC
      NP=NP1+NP2+NP3
      WRITE(6,121) N,NP,NP1,NP2,NP3,MI,NK,NGC,NGS,M,ICX,MS,DF,MSC
  121 FORMAT(/,' NO. OF OBSERVATIONS N:                ',I6,
     *       /,' TOTAL NO. OF PREDICTORS P:            ',I6,
     *       /,' NO. OF PREDICTORS/TIME SERIES         ',3I6,
     *       /,' MAX NO. OF INTERACTIONS MI:           ',I6,
     *       /,' MAX NO OF BASIS FUNCTIONS NK:         ',I6,
     *       /,' NO. OF RASTER POINTS FOR PLOTTING NGC:',I6,
     *       /,' NO. OF R.P. FOR SURFACE ESTIMATES NGS:',I6,
     *       /,' MODEL FLAG (1=LINEAR, 2=CUBIC)      M:',I6,
     *       /,' CONVEX HULL FLAG FOR PLOTS        ICX',I6,
```

```
     *       /,' MINIMUM SPAN:                          MS',I6,
     *       /,' DEGREES OF FREEDOM DF:                   ',F6.1,
     *       /,' MODE SELECTION CRITERIA:                 ',I6,
     *       /,'  (1=GCV, 2=AIC, 3=PC, 4=SC)    '       )
C
      IF(NP.LE.0) RETURN
      IPLX=1
      IPX=IPLX+NP
      IPY=IPX+N*NP
      IPW=IPY+N
      IPLAG=IPW+N
      IPIM=IPLAG+NP
      IPSP=IPIM + 21+NK*(3*MI+8)
C --- SP ALLOCATED FOR MARS OR PLOT, WHICHEVER LARGEST
      LSP=N*(MAX(NK,2)+3)+MAX(3*N+5*NK+NP,2*NP,4*N)+2*NP+4*NK
      LSP=MAX(LSP, 4*NGS*NGS, NGC, 2*N)
      IPMM=IPSP + LSP
      IF(IPMM .GT. NV) THEN
         WRITE(6,*) '**** MEMORY REQUIREMENTS EXCEEDED FOR X ******'
         WRITE(6,*) '**** MEMORY REQUESTED, AVAILABLE=',IPMM,NV
         STOP
      ENDIF
C
C --- READ LX, W AND LAGS
      CALL READLXW(INTV(IPLX),V(IPW), N,NP, INTV(IPLAG) )
C --- READ X AND Y
      CALL READXY(V(IPX),V(IPY),N,NP)
C --- COMPUTE NMCV,NTCV FROM THE DATA X
      CALL COMPCV(V(IPX),N,NP, INTV(IPLX), NMCV,NTCV, INTV(IPIM) )
C
      LMM=MAX(N*NP+2*MAX(MI,NMCV), 2*(MI+1), NMCV)
      IPFM=IPMM + LMM
      IPDP=IPFM + 3+NK*(5*MI+NMCV+6)+2*NP+NTCV
      IPDP=FLOAT(IPDP)/8. + 1.
      IPDP=8*IPDP
      IPEND=IPDP+8+2*MAX(N*NK,(NK+1)*(NK+1))+MAX((NK+2)*(NMCV+3),4*NK)
C*    WRITE(6,*) 'IP=LX,X,Y,W,IM,SP,MM,FM,DP,END=',
C*   *           IPLX,IPX,IPY,IPW,IPIM,IPSP,IPMM,IPFM,IPDP,IPEND
      IF(IPEND .GT. NV) THEN
         WRITE(6,*) '**** MEMORY REQUIREMENTS EXCEEDED ********'
         WRITE(6,*) '**** MEMORY REQUESTED, AVAILABLE=',IPEND,NV
         STOP
      ENDIF
```

```
C
      WRITE(6,*) 'LAGS AND LX FOR TIME SERIES 1'
      WRITE(6,122) (INTV(IPLAG-1+I),I=1,NP1)
      WRITE(6,122) (INTV(IPLX -1+I),I=1,NP1)
      WRITE(6,*) 'LAGS AND LX FOR TIME SERIES 2'
      WRITE(6,122) (INTV(IPLAG+NP1-1+I),I=1,NP2)
      WRITE(6,122) (INTV(IPLX +NP1-1+I),I=1,NP2)
      WRITE(6,*) 'LAGS AND LX FOR TIME SERIES 3'
      WRITE(6,122) (INTV(IPLAG+NP1+NP2-1+I),I=1,NP3)
      WRITE(6,122) (INTV(IPLX +NP1+NP2-1+I),I=1,NP3)
 122  FORMAT(20I5)
      WRITE(6,*) 'START MARS. MEMORY NEEDED/AVAILABLE=', IPEND,NV
      WRITE(6,*) '-------------------------------------------'
C
      CALL SETMS(MS)
      CALL SETDF(DF)
      CALL MARS(N,NP,V(IPX),V(IPY),V(IPW),NK,MI,V(IPLX),V(IPFM),
     *      V(IPIM),V(IPSP),V(IPDP),V(IPMM), MSC)
C
      CALL DISPFM(V(IPFM+1),V(IPDP),NK,MI, V(IPFM) )
C
C     CONSTRUCT PLOTS FOR INTERPRETING RESULTING MODEL:
C
      IF(NGC.EQ.0  .AND.  NGS.EQ.0) RETURN
C
      IPCRV=IPDP
      IPSRF=IPCRV +  2*NGC*NK
      IPEND=IPSRF + NGS*NGS*NK
      IF(IPEND .GT. NV) THEN
         WRITE(6,*) '**** MEMORY REQUIREMENTS EXCEEDED FOR PLOT *****'
         WRITE(6,*) '**** MEMORY REQUESTED, AVAILABLE=',IPEND,NV
         STOP
      ENDIF
      CALL PLOT (M,V(IPX),V(IPFM),V(IPIM),NGC,NGS,ICX,
     *        NC,V(IPCRV),NS,V(IPSRF),V(IPSP),V(IPMM) )
C WRITE PLOTS TO OUTPUT FILES FOR PLOTTING WITH LOCAL GRAPHICS PACKAGE:
C
      WRITE(6,*) 'PLOT =IPEND,NC,NS=',IPEND,NC,NS
      CALL WPLOT(V(IPCRV),V(IPSRF),NGC,NC, NGS,NS, FIN)
      RETURN
 999  CONTINUE
      WRITE(6,*) 'ERROR OPENING INPUT FILE:',FIN
      END
```

```
C
      SUBROUTINE READXY(X,Y,N,NP)
      REAL X(N,NP), Y(NP)
      DO 1 I=1,N
       READ(10,*) (X(I,J),J=1,NP),Y(I)
    1 CONTINUE
      END
C
      SUBROUTINE COMPCV(X,N,NP, LX, NMCV,NTCV, WK)
      REAL X(N,NP)
      INTEGER LX(NP)
      INTEGER WK(N)
      NMCV=0
      NTCV=0
C     FIND MAX OVER ALL COLUMNS AND SUM OF ALL COLUMN VALUES
      DO 30 J=1,NP
       IF(LX(J).LT.0) THEN
                NCAT=NUMCAT(X(1,J),N, WK)
       ELSE
                NCAT=0
       ENDIF
       NTCV=NTCV+NCAT
       NMCV=MAX(NMCV, NCAT)
   30 CONTINUE
      END
C
      FUNCTION NUMCAT(V,N, WK)
C --- FIND NUMBER OF DISTINCT VALUES IN V
      REAL V(N)
      INTEGER WK(N)
C     INITIALIZE DUPLIC. CONTROL MATRIX WK TO 0
      DO 10 I=1,N
       WK(I)=0
   10 CONTINUE
      NUMCAT=0
      DO 20 I=1,N
       VAL=V(I)
       IF(WK(I).EQ.0) THEN
                I1=I+1
                DO 30 J=I1,N
                 IF(VAL.EQ.V(J)) WK(J)=1
   30           CONTINUE
                NUMCAT=NUMCAT+1
```

```
       ENDIF
   20 CONTINUE
      END
C
      SUBROUTINE READLXW(LX,W, N,NP,LAG)
C --- READ VALUES FOR LX AND W
      INTEGER LX(NP),LAG(NP)
      REAL W(N)
      READ(10,*) LX
      READ(10,*) LAG
      READ(10,*) W
      END
C
      SUBROUTINE WPLOT(CRV,SRF,NGC,NC, NGS,NS, FNAME)
      REAL  CRV(NGC,2,NC), SRF(NGS,NGS,NS)
      CHARACTER*8 FNAME
      IF(NC.GT.0) THEN
           OPEN(11,FILE=FNAME // '.CUR',FORM='UNFORMATTED')
           WRITE(11) NGC,NC,CRV
      ENDIF
      IF(NS.GT.0) THEN
           OPEN(12,FILE=FNAME // '.SUR',FORM='UNFORMATTED')
           WRITE(12) NGS,NS,SRF
      ENDIF
      END
C
      SUBROUTINE DISPFM(FM,MMM,NK,MI,CONST)
C     ANALYSIS OF FM FOR MARS 3.0
      REAL FM(5,NK),MMM(NK,2,MI+1), VMIN(100),MMD1(100),MMD2(100)
      INTEGER OUT(100)
* WRITE OUT THE MODEL CONSTANT*
      WRITE(6,*) ' '
      WRITE(6,112) 'MODEL CONSTANT =     ',CONST
      WRITE(6,*) ' '
* INITIALIZE MMM
      DO 18 I=1,NK
       DO 17 J=1,2
        DO 16 K=1,MI+1
          MMM(I,J,K)=100.
   16   CONTINUE
   17  CONTINUE
   18 CONTINUE
C
```

```
      IC1=0
      DO 20 I=NK,1,-1
        IF(FM(1,I).EQ.0) GOTO 20
        IC1=IC1+1
        IC2=2
        MMM(IC1,2,1)=FM(1,I)
        MMM(IC1,1,IC2)=FM(2,I)
        MMM(IC1,2,IC2)=FM(3,I)
        DUM=FM(4,I)
   19   IF(DUM.EQ.0.0) GOTO 21
        IC2=IC2+1
        MMM(IC1,1,IC2)=FM(2,DUM)
        MMM(IC1,2,IC2)=FM(3,DUM)
        DUM=FM(4,DUM)
        GOTO 19
   21   MMM(IC1,1,1)=IC2-1
   20 CONTINUE
C
      DO 50  J=1,IC1
       KEND=MMM(J,1,1)
       DO 49  K=1,KEND
        MMD1(K)=ABS(MMM(J,1,K+1))
        VMIN(K)=MMM(J,1,K+1)
        MMD2(K)=MMM(J,2,K+1)
        OUT(K)=K
   49  CONTINUE
       CALL PSORT(MMD1,OUT,1,KEND)
       DO 48  K=1,KEND
        MMM(J,1,K+1)=VMIN(OUT(K))
        MMM(J,2,K+1)=MMD2(OUT(K))
   48  CONTINUE
   50 CONTINUE
C
C --- PRINT INTERACTION VARIABLE REPORT
      DO 100 I=1,MI
       WRITE(6,*) '************* '
       WRITE(6,*) 'INTERACTION LEVEL:',I
       WRITE(6,*) '************* '
  101  CONTINUE
       DO 102 J=1,MI
        VMIN(J)=999999.
  102  CONTINUE
       DO 110 J=1,IC1
```

```
      KEND=MMM(J,1,1)+1
      IF(KEND-1 .NE. I) GO TO 110
      DO 120 K=2, KEND
       IF(ABS(MMM(J,1,K)).LT.VMIN(K-1)) THEN
                J1=J
                DO 130 KK=2, KEND
                 VMIN(KK-1)=ABS(MMM(J,1,KK))
130             CONTINUE
                GO TO 110
       ELSE
           IF(ABS(MMM(J,1,K)).GT.VMIN(K-1)) GO TO 110
       ENDIF
120    CONTINUE
110   CONTINUE
C
      IF(VMIN(1).NE.999999.) THEN
           WRITE(6,111) '      VARIABLES     ',(MMM(J1,1,KK),KK=2,I+1)
           WRITE(6,112) 'COEFF AND KNOTS     ',(MMM(J1,2,KK),KK=1,I+1)
           WRITE(6,*) '  '
           MMM(J1,1,1)=MMM(J1,1,1) + MI
           GO TO 101
      ENDIF
100  CONTINUE
     RETURN
111  FORMAT(A25,15X,7F10.3)
112  FORMAT(A25,F15.6,7F10.3)
     END
```

APPENDIX D. MARS OUTPUT FOR ASTAR MODEL GRANITE2

```
 DRIVER FOR MARS 3.X.

NO. OF OBSERVATIONS N:                    4330
TOTAL NO. OF PREDICTORS P:                  52
NO. OF PREDICTORS/TIME SERIES               50     1     1
MAX NO. OF INTERACTIONS MI:                  3
MAX NO OF BASIS FUNCTIONS NK:               60
NO. OF RASTER POINTS FOR PLOTTING NGC:       0
NO. OF R.P. FOR SURFACE ESTIMATES NGS:       0
MODEL FLAG (1=LINEAR, 2=CUBIC)       M:      0
CONVEX HULL FLAG FOR PLOTS          ICX      0
MINIMUM SPAN:                        MS     50
DEGREES OF FREEDOM DF:                     3.0
MODEL SELECTION CRITERIA (MSC):            GCV
 (1=GCV, 2=AIC, 3=PC, 4=SC)
LAGS AND LX FOR TIME SERIES 1
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
  16   17   18   19   20   21   22   23   24   25   26   27   28   29   30
  31   32   33   34   35   36   37   38   39   40   41   42   43   44   45
  46   47   48   49   50
   1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
   1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
   1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
   1    1    1    1    1
LAGS AND LX FOR TIME SERIES 2
   0
   3
LAGS AND LX FOR TIME SERIES 3
   0
   3
START MARS. MEMORY NEEDED/AVAILABLE=     1150368     1200000
-------------------------------------------


MARS MODELING, VERSION 3.5a (6/16/91)


INPUT PARAMETERS (SEE DOC.):
```

N	P	NK	MS	MI	DF	IL	FV	IC
4330	52	60	50	3	3.000	0	0.000	0

PREDICTOR VARIABLE FLAGS:

VAR:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	
41	42	43	44	45	46	47	48	49	50	51	52									
FLAG:	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		
1	1	1	1	1	1	1	1	1	1	3	3									

ORDERABLE RESPONSE:

MIN	N/4	N/2	3N/4	MAX
8.000	10.60	11.70	12.90	17.00

THERE ARE 52 ORDERABLE PREDICTOR VARIABLES.

VAR	MIN	N/4	N/2	3N/4	MAX
1	8.000	10.60	11.70	12.90	17.00
2	8.000	10.60	11.70	12.90	17.00
3	8.000	10.60	11.70	12.90	17.00
4	8.000	10.60	11.70	12.90	17.00
5	8.000	10.50	11.70	12.90	17.00
6	8.000	10.50	11.70	12.90	17.00
7	8.000	10.50	11.70	12.90	17.00
8	8.000	10.50	11.70	12.90	17.00
9	8.000	10.50	11.70	12.90	17.00
10	8.000	10.50	11.70	12.90	17.00
11	8.000	10.50	11.70	12.90	17.00
12	8.000	10.50	11.70	12.90	17.00
13	8.000	10.50	11.60	12.90	17.00
14	8.000	10.50	11.60	12.90	17.00
15	8.000	10.50	11.60	12.90	17.00
16	8.000	10.50	11.60	12.90	17.00
17	8.000	10.50	11.60	12.90	17.00
18	8.000	10.50	11.60	12.90	17.00
19	8.000	10.50	11.60	12.90	17.00
20	8.000	10.50	11.60	12.90	17.00
21	8.000	10.50	11.60	12.90	17.00
22	8.000	10.50	11.60	12.90	17.00
23	8.000	10.50	11.60	12.90	17.00
24	8.000	10.50	11.60	12.90	17.00
25	8.000	10.50	11.60	12.90	17.00

26	8.000	10.50	11.60	12.90	17.00
27	8.000	10.50	11.60	12.90	17.00
28	8.000	10.50	11.60	12.90	17.00
29	8.000	10.50	11.60	12.90	17.00
30	8.000	10.50	11.60	12.90	17.00
31	8.000	10.50	11.60	12.90	17.00
32	8.000	10.50	11.60	12.90	17.00
33	8.000	10.50	11.60	12.90	17.00
34	8.000	10.50	11.60	12.90	17.00
35	8.000	10.50	11.60	12.80	17.00
36	8.000	10.50	11.60	12.80	17.00
37	8.000	10.50	11.60	12.80	17.00
38	8.000	10.50	11.60	12.80	17.00
39	8.000	10.50	11.60	12.80	17.00
40	8.000	10.50	11.60	12.80	17.00
41	8.000	10.50	11.60	12.80	17.00
42	8.000	10.50	11.60	12.80	17.00
43	8.000	10.50	11.60	12.80	17.00
44	8.000	10.50	11.60	12.80	17.00
45	8.000	10.50	11.60	12.80	17.00
46	8.000	10.50	11.60	12.80	17.00
47	8.000	10.50	11.60	12.80	17.00
48	8.000	10.50	11.60	12.80	17.00
49	8.000	10.50	11.60	12.80	17.00
50	8.000	10.50	11.60	12.80	17.00
51	-1.000	-0.7000	0.0000E+00	0.7000	1.000
52	-1.000	-0.7000	0.0000E+00	0.7000	1.000

FORWARD STEPWISE KNOT PLACEMENT:

BASFN(S)		MSC	#INDBSFNS	#EFPRMS	VARIABLE	KNOT	PARENT
0		2.6000	0.0	1.0			
2	1	0.3031	2.0	5.9	1.	15.40	0.
3		0.2983	3.0	9.8	14.	8.000	0.
4		0.2972	4.0	13.8	2.	8.000	0.
6	5	0.2959	6.0	18.7	19.	9.100	4.
7		0.2944	7.0	22.6	51.	-1.000	0.
9	8	0.2928	9.0	27.5	35.	10.00	5.
11	10	0.2923	11.0	32.5	17.	13.40	4.
13	12	0.2918	13.0	37.4	2.	14.90	2.
15	14	0.2915	15.0	42.3	3.	14.80	3.
16		0.2905	16.0	46.2	2.	8.000	14.
18	17	0.2901	18.0	51.2	2.	14.80	3.

19	0.2892	19.0	55.1	36.	8.000	17.
21 20	0.2886	21.0	60.0	1.	14.90	11.
23 22	0.2883	23.0	64.9	7.	11.80	11.
25 24	0.2880	25.0	69.8	1.	13.10	18.
27 26	0.2878	27.0	74.8	31.	13.40	10.
28	0.2874	28.0	78.7	26.	8.000	17.
30 29	0.2873	30.0	83.6	39.	15.00	18.
32 31	0.2871	32.0	88.5	36.	12.40	4.
34 33	0.2864	34.0	93.5	1.	14.30	32.
36 35	0.2862	36.0	98.4	3.	13.60	4.
38 37	0.2861	38.0	123.3	47.	11.80	31.
40 39	0.2860	40.0	138.2	35.	13.30	36.
42 41	0.2858	42.0	133.2	5.	12.80	31.
44 43	0.2857	44.0	148.1	1.	10.90	31.
46 45	0.2855	46.0	123.0	29.	10.10	15.
48 47	0.2852	48.0	137.9	20.	9.500	0.
50 49	0.2852	50.0	132.8	45.	13.00	35.
52 51	0.2851	52.0	147.8	44.	14.90	10.
54 53	0.2851	54.0	122.7	30.	10.20	32.
56 55	0.2849	56.0	137.6	35.	15.40	3.
58 57	0.2849	58.0	132.5	15.	15.00	13.
60 59	0.2848	60.0	147.5	25.	9.500	56.

FINAL MODEL AFTER BACKWARD STEPWISE ELIMINATION:

BSFN:	0	1	2	3	4	5
COEF:	15.778	1.2432	-1.0419E+00	0.0000E+00	-0.7511E-01	-0.1441E-01

BSFN:	6	7	8	9	10	11
COEF:	-0.1587	-0.1026	0.4741E-02	-0.2700E-01	-0.2597E-01	-0.1784E-01

BSFN:	12	13	14	15	16	17
COEF:	0.3677E+00	0.0000E+00	-0.3553E+00	0.0000E+00	0.4409E-01	-0.2146E+00

BSFN:	18	19	20	21	22	23
COEF:	0.0000E+00	0.5120E-01	0.5373E-01	0.5596E-02	-0.7694E-02	-0.5888E-02

BSFN:	24	25	26	27	28	29
COEF:	-0.1822E-01	0.0000E+00	0.1478E-01	0.1307E-01	-0.3072E-01	0.4366E-01

BSFN:	30	31	32	33	34	35
COEF:	0.0000E+00	0.0000E+00	-0.2142E-01	-0.5539E-01	0.0000E+00	0.0000E+00

```
BSFN:      36          37          38          39          40          41
COEF:  0.0000E+00  0.0000E+00 -0.2932E-01  0.2563E-01  0.0000E+00  0.0000E+00

BSFN:      42          43          44          45          46          47
COEF: -0.2426E-01 -0.3663E-02  0.7473E-01 -0.2695E-02 -0.1881E-01  0.5118E-01

BSFN:      48          49          50          51          52          53
COEF:  0.0000E+00  0.1456E-01  0.1286E-01 -0.4244E-01  0.0000E+00  0.5698E-02

BSFN:      54          55          56          57          58          59
COEF:  0.1640E-01 -0.4912E-01  0.1513E-01  0.5708E-01  0.1159E-02  0.0000E+00

BSFN:      60
COEF:  0.1651E-01

  (PIECEWISE LINEAR) MSC =    0.2808      #EFPRMS = 115.7

ANOVA DECOMPOSITION ON 44 BASIS FUNCTIONS:
 FUN. STD. DEV.     -MSC    #BSFNS  #EFPRMS  VARIABLE(S)
  1    1.682      0.4456       2       5.8     1
  2   0.1211      0.2812       1       2.6     2
  3   0.7252E-01  0.2846       1       2.6    51
  4   0.7980E-01  0.2814       1       2.6    20
  5   0.1451      0.2813       2       5.2     2  19
  6   0.7697E-01  0.2817       2       2.9     2  17
  7   0.4403E-01  0.2815       1       5.8     1   2
  8   0.3416      0.2820       1       2.9     3  14
  9   0.1995      0.2813       1       2.9     2  14
 10   0.7316E-01  0.2819       1       5.8     2  36
 11   0.9518E-01  0.2828       2       2.9    14  35
 12   0.2111      0.2825       2       2.9     2  19  35
 13   0.3274      0.2817       1       5.8     2   3  14
 14   0.2728      0.2822       1       2.9     2  14  36
 15   0.8892E-01  0.2816       2       2.9     1   2  17
 16   0.4887E-01  0.2814       2       5.8     2   7  17
 17   0.3807E-01  0.2814       1       2.9     1   2  14
 18   0.5182E-01  0.2810       2       2.9     2  17  31
 19   0.1540      0.2813       1       5.8     2  14  26
 20   0.3989E-01  0.2819       1       2.9     2  14  39
 21   0.7842E-01  0.2822       3       2.9     1   2  36
 22   0.2995E-01  0.2813       1       5.8     2  36  47
 23   0.4398E-01  0.2816       1       2.9     2   3  35
 24   0.4913E-01  0.2818       1       2.9     2   5  36
```

```
 25   0.6054E-01  0.2822       2       5.8      3  14  29
 26   0.6108E-01  0.2816       2       2.9      2   3  45
 27   0.3383E-01  0.2814       1       2.9      2  17  44
 28   0.4288E-01  0.2809       2       5.8      2  30  36
 29   0.6748E-01  0.2813       2       2.9      1   2  15
 30   0.2573E-01  0.2810       1       2.9     14  25  35

PIECEWISE CUBIC FIT ON 44 BASIS FUNCTIONS, MSC =  .2867

-MSC FOR REMOVING EACH VARIABLE:
     0.5545      0.2981      0.2838      0.2808      0.2818
     0.2808      0.2814      0.2808      0.2808      0.2808
     0.2808      0.2808      0.2808      0.2876      0.2813
     0.2808      0.2832      0.2808      0.2822      0.2814
     0.2808      0.2808      0.2808      0.2808      0.2810
     0.2813      0.2808      0.2808      0.2823      0.2809
     0.2810      0.2808      0.2808      0.2808      0.2837
     0.2868      0.2808      0.2808      0.2819      0.2808
     0.2808      0.2808      0.2808      0.2814      0.2816
     0.2808      0.2813      0.2808      0.2808      0.2808
     0.2846      0.2808

RELATIVE VARIABLE IMPORTANCE:
      100.0       25.11       10.40      0.0000E+00   6.115
     0.0000E+00   4.666      0.0000E+00  0.0000E+00  0.0000E+00
     0.0000E+00  0.0000E+00  0.0000E+00   15.77       4.220
     0.0000E+00   9.337      0.0000E+00   7.103       4.619
     0.0000E+00  0.0000E+00  0.0000E+00  0.0000E+00   2.542
      4.061      0.0000E+00  0.0000E+00   7.224       1.871
      2.629      0.0000E+00  0.0000E+00  0.0000E+00   10.26
      14.83      0.0000E+00  0.0000E+00   6.397      0.0000E+00
     0.0000E+00  0.0000E+00  0.0000E+00   4.539       5.375
     0.0000E+00   4.024      0.0000E+00  0.0000E+00  0.0000E+00
      11.75      0.0000e+00

-------------------- MARS OUTPUT MODEL --------------

THE VARIABLE SIGN INDICATES A LEFT (- SIGN) OR RIGHT ( SIGN)
TRUNCATED SPLINE FUNCTION WITH THE INDICATED KNOT

    MODEL CONSTANT =          15.778527

*************
```

```
INTERACTION LEVEL:          1
*************
          VARIABLES                    -1.000
    COEFF AND KNOTS          -1.041947   15.400

          VARIABLES                     1.000
    COEFF AND KNOTS           1.124321   15.400

          VARIABLES                     2.000
    COEFF AND KNOTS          -0.075109    8.000

          VARIABLES                    20.000
    COEFF AND KNOTS           0.051177    9.500

          VARIABLES                    51.000
    COEFF AND KNOTS          -0.102553   -1.000

*************
INTERACTION LEVEL:          2
*************
          VARIABLES                    -1.000    2.000
    COEFF AND KNOTS           0.367790   15.400   14.900

          VARIABLES                     2.000   14.000
    COEFF AND KNOTS          -0.214606   14.800    8.000

          VARIABLES                     2.000  -17.000
    COEFF AND KNOTS          -0.017842    8.000   13.400

          VARIABLES                     2.000   17.000
    COEFF AND KNOTS          -0.025969    8.000   13.400

          VARIABLES                     2.000  -19.000
    COEFF AND KNOTS          -0.158722    8.000    9.100

          VARIABLES                     2.000   19.000
    COEFF AND KNOTS          -0.014415    8.000    9.100

          VARIABLES                     2.000  -36.000
    COEFF AND KNOTS          -0.021417    8.000   12.400

          VARIABLES                     3.000   14.000
    COEFF AND KNOTS          -0.355389   14.800    8.000
```

```
          VARIABLES                      14.000   -35.000
    COEFF AND KNOTS          0.015131     8.000    15.400

          VARIABLES                      14.000    35.000
    COEFF AND KNOTS         -0.049118     8.000    15.400

*************
INTERACTION LEVEL:          3
*************
          VARIABLES                       1.000    -2.000    14.000
    COEFF AND KNOTS         -0.018219    13.100    14.800     8.000

          VARIABLES                      -1.000    -2.000   -15.000
    COEFF AND KNOTS          0.001159    15.400    14.900    15.000

          VARIABLES                      -1.000    -2.000    15.000
    COEFF AND KNOTS          0.057077    15.400    14.900    15.000

          VARIABLES                      -1.000     2.000   -17.000
    COEFF AND KNOTS          0.005596    14.900     8.000    13.400

          VARIABLES                       1.000     2.000    17.000
    COEFF AND KNOTS          0.053733    14.900     8.000    13.400

          VARIABLES                      -1.000     2.000    36.000
    COEFF AND KNOTS          0.074733    10.900     8.000    12.400

          VARIABLES                       1.000     2.000    36.000
    COEFF AND KNOTS         -0.003663    10.900     8.000    12.400

          VARIABLES                       1.000     2.000   -36.000
    COEFF AND KNOTS         -0.055390    14.300     8.000    12.400

          VARIABLES                       2.000     3.000    14.000
    COEFF AND KNOTS          0.044091     8.000    14.800     8.000

          VARIABLES                       2.000    -3.000    35.000
    COEFF AND KNOTS          0.025631     8.000    13.600    13.300

          VARIABLES                       2.000     3.000   -45.000
    COEFF AND KNOTS          0.012861     8.000    13.600    13.000
```

VARIABLES		2.000	3.000	45.000
COEFF AND KNOTS	0.014562	8.000	13.600	13.000
VARIABLES		2.000	-5.000	36.000
COEFF AND KNOTS	-0.024261	8.000	12.800	12.400
VARIABLES		2.000	-7.000	-17.000
COEFF AND KNOTS	-0.005888	8.000	11.800	13.400
VARIABLES		2.000	7.000	-17.000
COEFF AND KNOTS	-0.007694	8.000	11.800	13.400
VARIABLES		2.000	14.000	26.000
COEFF AND KNOTS	-0.030715	8.000	8 000	8.000
VARIABLES		2.000	14.000	36.000
COEFF AND KNOTS	0.051202	8.000	8.000	8.000
VARIABLES		2.000	14.000	39.000
COEFF AND KNOTS	0.043657	8.000	8.000	15.000
VARIABLES		2.000	17.000	-31.000
COEFF AND KNOTS	0.013065	8.000	13.400	13.400
VARIABLES		2.000	17.000	31.000
COEFF AND KNOTS	0.014785	8.000	13.400	13.400
VARIABLES		2.000	17.000	44.000
COEFF AND KNOTS	-0.042443	8.000	13.400	14.900
VARIABLES		2.000	19.000	-35.000
COEFF AND KNOTS	-0.027002	8.000	9.100	10.000
VARIABLES		2.000	19.000	35.000
COEFF AND KNOTS	0.004741	8.000	9.100	10.000
VARIABLES		2.000	-30.000	-36.000
COEFF AND KNOTS	0.016398	8.000	10.200	12.400
VARIABLES		2.000	30.000	-36.000
COEFF AND KNOTS	0.005698	8.000	10.200	12.400
VARIABLES		2.000	36.000	-47.000

COEFF AND KNOTS	-0.029326	8.000	12.400	11.800
VARIABLES		-3.000	14.000	-29.000
COEFF AND KNOTS	-0.018812	14.800	8.000	10.100
VARIABLES		-3.000	14.000	29.000
COEFF AND KNOTS	-0.002695	14.800	8.000	10.100
VARIABLES		14.000	-25.000	-35.000
COEFF AND KNOTS	0.016511	8.000	9.500	15.400

LIST OF REFERENCES

Akaike, H., "A New Look at the Statistical Identification Model", *IEEE*, v. AC-19, pp. 716–723, 1974.

Akaike, H., "A Bayesian Analysis of the Minimum AIC Procedure", *Ann Inst. Stat. Math*, v. 30, pp. 9–14, 1979.

Altman, N. S., *Smoothing Data with Correlated Errors*, Department of Statistics, Stanford University, Report 280, October 1987.

Amemiya, T., "Selection of Regressors", *International Economic Review*, v. 21, pp. 331–354, 1980.

Bellman, R. E., *Adaptive Control Processes*, Princeton University Press, Princeton, New York, 1961.

Box, G. and Tiao, G. C., "The Canonical Analysis of Multiple Time Series", *Biometrika*, v. 64, pp. 355–365, 1977.

Breaker, L. C. and Lewis, P. A. W., *On the Detection of a 40 to 50 Day Oscillation in Sea-Surface Temperature Along the Central California Coast*, Department of Operations Research, Naval Postgraduate School, Report NPS55-85-025, September 1985.

Breaker, L. C. and Lewis, P. A. W., "A 40-50 Day Oscillation in Sea-Surface Temperature Along the Central California Coast", *Estuarine, Coastal and Shelf Science*, v. 26, pp. 395–408, 1988.

Breiman, L., Friedman, J., Olshen, R., and Stone, C., *Classification and Regression*

Trees, Wadsworth, Belmont, CA., 1984.

Breiman, L. and Meisel, W. S., "General Estimates of the Intrinsic Variability of Data in Nonlinear Regression Models", *Journal of the American Statistical Association*, v. 71, pp. 301–307, 1976.

Cleveland, W. S., "Robust Locally Weighted Regression and Smoothing Scatterplots", *Journal of the American Statistical Association*, v. 74, no. 368, pp. 829–836, December 1979.

Craven, P. and Wahba, G., "Smoothing Noisy Data With Spline Functions. Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation", *Numerische Mathematik*, v. 31, pp. 317–403, 1979.

Draper, N. and Smith, H., *Applied Regression Analysis*, Wiley, New York, 1966.

Eubank, R. L., *Spline Smoothing and Nonparametric Regression*, Marcel Dekker Inc., New York, 1988.

Friedman, J. H., *A Variable Span Smoother*, Department of Statistics, Stanford University, Report 5, November 1984.

Friedman, J. H., *Multivariate Adaptive Regression Splines*, Department of Statistics, Stanford University, Report 102, November 1988.

Friedman, J. H., "Multivariate Adaptive Regression Splines", *Annals of Statistics*, v. 19, no. 2, pp. 1–142, 1991.

Gelb, A., *Applied Optimal Estimation*, M.I.T. Press, Cambridge, Mass, 1974.

Granger, C. W. and Anderson, A. P., "An Introduction to Bilinear Time Series Models", *Vandenhoeck and Ruprecht*, pp. 1–94, 1978.

Gudmundsson, G., "Short Term Variations of a Glacier-Fed River", *Tellus XXII*, v. 3, pp. 341–353, 1970.

Izenman, A. J., "J. R. Wolf and J. A. Wolfer: An Historical Note on the Zurich Sunspot Relative Numbers", *Journal of the Royal Statistical Society, (Ser A)*, v. 146, no. 3, pp. 311–318, 1983.

Judge, G., Hill, R., Griffiths, W., Lutkepohl, H., and Lee, T., *Introduction to the Theory and Practice of Econometrics*, 2th ed., Wiley, 1985.

Kendall, M., Stuart, A., and Ord, J. K., *The Advanced Theory of Statistics*, 4th ed., v. 3, Macmillan, 1983.

Lawrance, A. J. and Kottegoda, N. T., "Stochastic Modelling of Riverflow Time Series", *Journal of the Royal Statistical Society, (Ser A)*, v. 140, no. 1, pp. 1–47, 1977.

Lewis, P. A. W. and Orav, J., *Simulation Methodology for Statisticians, Operations Analysts, and Engineers*, v. 1, Wadsworth & Brooks/Cole, 1988.

Lewis, P. A. W. and Stevens, J. G., "Smoothing Time Series for Input and Output Analysis in System Simulation Experiments", In *Proceedings of the 1990 Winter Simulation Conference*, Balci, Sadowski, Nance, O., P., R., and E., R., editors, pp. 48–50, 1990.

Lin, C. and Mudhoekar, G. S., "A Simple Test for Normality Against Asymmetric Alternatives", *Biometrika*, v. 67, pp. 455–461, 1980.

Mallows, C. L., "Some Comments on C_p", *Technometrics*, v. 15, pp. 661–675, 1973.

Moeanaddin, R., *Aspects of Non-linear Time Series Analysis*, PhD thesis, University of Kent, England, 1989.

Morgan, J. N. and Sonquist, J. A., "Problems in the Analysis of Survey Data, and a

Proposal", *Journal of the American Statistical Association*, v. 58, pp. 415–434, 1963.

Parzen, E., "Some Recent Advances in Time Series Modeling", *IEEE*, v. AC-19, no. 6, pp. 723–730, December 1974.

Priestley, M. B., *Spectral Analysis and Time Series*, Academic Press, 1981.

Priestley, M. B., *Non-Linear and Non-Stationary Times Series*, Academic Press, 1988.

Rao, P., "Some Notes on Misspecification in Multiple Regression", *The American Statistician*, v. 25, pp. 37–39, 1971.

Rao, T. S. and Gabr, M. M., *An Introduction to Bispectral Analysis and Bilinear Time Series Models*, Springer-Verlag, 1984.

Rissanen, J., "Modeling by Shortest Data Description", *Automatica*, v. 14, pp. 465–471, 1978.

Rissanen, J., "Stochastic Complexity", *Journal of the Royal Statistical Society (B)*, v. 49, no. 3, pp. 223–239, 1987.

Schwarz, G., "Estimating the Dimension of a Model", *Annals of Statistics*, v. 6, pp. 461–464, 1978.

Shumaker, L. L., *Spline Functions*, John Wiley and Sons, Inc., 1981.

Silverman, B. W., "Some Aspects of the Spline Smoothing Approach to Non-Parametric Regression Curve Fitting", *Journal of the Royal Statistical Society (Ser B)*, v. 47, no. 1, pp. 1–52, 1985.

Smith, P. L., "Splines as a Useful and Convenient Statistical Tool", *The American Statistician*, v. 33, no. 2, pp. 57–62, 1979.

Stone, C. J., "Consistent Nonparametric Regression", *The Annals of Statistics*, v. 5, no. 4, pp. 595–645, 1977.

Thisted, R. A., *Elements of Statistical Computing: Numerical Computation*, Chapman and Hall, 1988.

Tiao, G. C. and Tsay, R. S., "Model Specification in Multivariate Time Series", *Journal of the Royal Statistical Association (Ser B)*, v. 51, no. 2, pp. 157–213, 1989.

Tong, H., *Threshold Models in Non-linear Time Series Analysis*, Springer-Verlag, 1983.

Tong, H., *Nonlinear Time Series*, Oxford University Press, 1990.

Tong, H. and Lim, K. S., "Threshold Autoregression, Limit Cycles and Cyclical Data", *Journal of the Royal Statistical Society (Ser B)*, v. 42, no. 3, pp. 245–292, 1980.

Tong, H., Thanoon, B., and Gudmundsson, G., "Threshold Time Series Modeling of Two Icelandic Riverflow Systems", *Water Resources Bulletin*, v. 21, no. 4, pp. 651–660, 1985.

Tsay, R. S., "Testing and Modeling Threshold Autoregressive Processes", *Journal of the American Statistical Association*, v. 84, no. 405, pp. 231–240, 1989.

Wegman, E. J. and Wright, I. W., "Splines in Statistics", *Journal of the American Statistical Association*, v. 78, no. 382, pp. 351–365, 1983.

Wold, S., "Spline Functions in Data Analysis", *Technometrics*, v. 16, no. 1, pp. 1–11, February 1974.

Yule, G. U., "On a Method of Investigating Periodicities in Disturbed Series with Special Reference to Wolfer's Sunspot Numbers", *Philos. Trans. Roy. Soc (Ser A)*, v. 226, pp. 267–298, 1927.

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center — 2
   Cameron Station
   Alexandria, Virginia 22304-6145

2. Library, Code 52 — 2
   Naval Postgraduate School
   Monterey, California 93943-5002

3. Professor Peter A. W. Lewis, Code ORLw — 2
   Naval Postgraduate School
   Monterey, California 93943-5000

4. Professor Peter Purdue, Code ORPd — 1
   Naval Postgraduate School
   Monterey, California 93943-5000

5. Professor Charles W. Therrien, Code ECTi — 1
   Naval Postgraduate School
   Monterey, California 93943-5000

6. Professor Maurice Weir, Code MAWc — 1
   Naval Postgraduate School
   Monterey, California 93943-5000

7. Professor Lyn R. Whitaker, Code ORWh — 1
   Naval Postgraduate School
   Monterey, California 93943-5000

8. MAJ James G. Stevens, — 2
   2634 Childs Lane
   Alexandria, Virginia 22308